Processing of natural sounds and scenes in the human brain

Processing of natural sounds and scenes in the human brain

Noël Staeren

ISBN 978 94 6159 311 5
Datawyse / Universitaire Pers Maastricht

The work in this thesis was supported by the The Netherlands Organisation for Scientific Research (NWO) and was conducted at Maastricht University, the Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, and the Low Temperature Laboratory, University of Technology, Helsinki.

Processing of natural sounds and scenes in the human brain

DISSERTATION

to obtain the degree of Doctor at
Maastricht University
on the authority of the Rector Magnificus, Prof. dr. L.L.G. Soete
in accordance with the decision of the Board of Deans
to be defended on
Friday 21st of March 2014, at 10:00 hours

door

Noël Paul Marie Clement Staeren

Geboren op 6 januari 1980 te Bilzen

UNIVERSITAIRE PERS MAASTRICHT

Supervisor

Prof. dr. E. Formisano

Co-supervisors

Prof. dr. R. W. Goebel

Dr. H. Renvall

Assessment Committee

Prof. dr. A. T. Sack (Chairman)

Dr. M. L. Bonte

Prof. dr. F. Di Salle (University of Salerno, Italy)

Prof. dr. B. Jansma

Contents

CHAPTER 1

General introduction

Introduction

The research described in this thesis investigates the relationship between human brain activity and the perception of natural sounds. Most experimental studies in the field of auditory neuroscience use synthetic sounds. These sounds are useful because they allow experimenters a great level of control over their physical parameters, which makes them most suitable for investigating the neural processing of basic acoustic features. However, for studying the auditory system "in action", more complex and ecologically valid sounds may be more appropriate because they engage the cortex and the brain in meaningful processing. The research reported in this thesis uses natural sounds in combination with functional brain imaging to examine two relevant aspects of audition. The first aspect relates to the ability of humans and animals to recognize sounds in natural environments. What are the cortical mechanisms enabling this sound recognition? Does the brain have specific representations of natural sound categories? How do these putative representations relate to the physical properties of the sounds? These research questions are addressed experimentally using functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) in combination with advanced data analysis techniques (chapters 2 and 3).

The second part of the thesis deals with the so-called 'auditory scene analysis' problem. Auditory scene analysis refers to the processes required for deriving descriptions of individual sources ('auditory objects' or 'auditory streams') from mixtures of simultaneous sounds (Bregman, 1990). Because natural environments typically involve multiple sound sources, auditory scene analysis represents a crucial aspect of human (and animal) hearing, which lies at the heart of the ability to select and respond to relevant acoustic stimuli even when these are masked by competing sound sources or background noise. How are sounds extracted from the mixture of other overlapping sounds? Does the auditory cortex use spatial information for segregating overlapping sounds? How does the representation of a sound in a mixture relate to the representation of the same sound presented against a silent background? These questions are considered in the second part of the thesis, which includes two fMRI studies employing respectively binaural natural scenes and musical recordings as stimuli (chapters 4 and 5).

The human auditory system

The auditory system translates the acoustic input at the ears into the experience of hearing. Sound waves that enter the ear are first filtered in the outer ear and middle ear before they are transmitted further to the inner ear (cochlea). The cochlea contains hair cells that translate the physical motion produced by the sound wave into electrochemical signals. Each cell responds maximally to a specific part of the auditory frequency spectrum, the so-called best frequency of that cell. The hair cells are arranged across the longitudinal axis of the cochlea such that cells with neighboring best frequencies are located adjacently. This arrangement is

called tonotopy and is preserved along the entire auditory pathway, which includes amongst others, the thalamus and the primary auditory cortex (Lee et al., 2004). The human primary auditory cortex, the first cortical stage of auditory processing, is located on Heschl's gyrus (HG) and in some cases, it extends to the adjacent Heschl's sulcus (HS) (Formisano et al., 2003; Hackett et al., 1998).

Even though acoustic input has already passed several neural processing stages before reaching the cortex (Anderson et al., 2009), it is assumed that it is cortical processing that is most relevant for the extraction of higher-order sound attributes and the representation of sound percepts (Bizley et al., 2009; Formisano et al., 2008a; Griffiths, 2003; Kaas and Hackett, 1999; Lewis et al., 2005; Rauschecker and Tian, 2000; Riecke et al., 2007; Staeren et al., 2009). Beyond the primary auditory cortex, it has been suggested that sound processing related to the recognition of sounds proceeds along the anterior/ventral "what" stream, whereas processing of spatial information for sound localization proceeds along the posterior/dorsal "where" stream (Alain et al., 2001; Romanski et al., 1999).

One of the goals of this thesis is to understand the cortical representation and processing of natural sounds and scenes within these processing streams.

Measuring brain activation

The experimental studies presented in this thesis use fMRI and MEG to measure brain activation during the perception of natural sounds and scenes. FMRI is a measurement technique based on magnetic resonance (MR) that relies on the detection of stimulus-evoked changes of blood-oxygenation-level-dependent (BOLD) contrast in the vascular (hemodynamic) system that supplies blood to the neuronal tissue. The BOLD response is an endogenous MR contrast that is measured because of the different magnetic properties of the blood in its oxygenated and deoxygenated states (Ogawa et al., 1993). Changes of BOLD contrast have been shown to be closely related to neuronal activity (Logothetis et al., 2001) and fMRI has been used in many studies of brain function. Even though the spatial resolution of fMRI can be relatively high (e.g. 1 mm^3 is feasible in most high field scanners), its temporal resolution is rather poor (> hundreds of milliseconds). This sluggishness occurs because the BOLD response takes more time to evolve compared to the neuronal activity.

MEG detects small magnetic fields that are generated by the net activity of neuronal populations. If neural currents synchronize, for example when listening to a sound, a small but detectable magnetic field is generated. The MEG scanner uses arrays of superconducting quantum interference devices (SQUIDs) which are able to detect these very small magnetic fields from the scalp of the subject (Cohen, 1972). Because the MEG signal provides a relatively direct measure of the electromagnetic properties of neural activity, it can resolve the timing of this neural activity with higher precision than methods based on neurovascular coupling, such as fMRI. However because MEG is recorded on the scalp, the neural sources underlying the recorded signals cannot be easily localized and need to be estimated based

on sophisticated models (Hari et al., 2010), which significantly decreases the spatial resolution of the measurements.

Analysis methods

FMRI studies conventionally use experimental setups where specific events, such as stimuli or behavior (e.g., button responses) occur sequentially and fMRI measurements are done simultaneously. The events and measurements are repeated several times to increase the number of samples obtained per experimental condition, which increases the power of statistical tests applied to the obtained data. FMRI data analysis typically begins by modeling the expected BOLD responses to the experimental events using a general linear model (GLM). To that end, the time courses of the different events are convolved with a function that describes the expected shape of the hemodynamic response (HRF) (Friston et al., 1995). The modeled BOLD responses are then fitted to the obtained BOLD responses using least squares regression. The resulting regression weights are taken as an estimate of the brain's response to the different events. These analyses are done separately for each voxel. Each voxel represents a local fMRI measurement from a different small subvolume of the imaged tissue. Statistical tests are then applied to the regression weights in order to assess whether brain responses to events resembling different experimental conditions differ significantly.

MEG data analysis involves similar procedures. However, additional assumptions need to be made in order to estimate the location of the neural sources that produce the activity measured at the scalp. One way to do this is using equivalent current dipoles (ECDs) (Hämäläinen et al., 1993). An ECD represents the hypothesized location, orientation, and strength of a net current in an activated brain region. Typical source analyses focus on only a subset of ECDs – those that can explain more than some fixed percentage (e.g. 85%, see chapter 3) of the variance in the local magnetic field that is obtained at the scalp during the response peak. The head of the subject is usually modeled as a homogeneous sphere or using more complex shapes derived from anatomical images of the head obtained with MR imaging (see chapter 3).

As described above, most standard fMRI data analyses are conducted separately for each voxel. A limitation of such univariate analyses is that they do not take into account correlations between different voxels. Each voxel is therefore characterized separately from the others. In contrast, multivariate analyses exploit the correlations between different voxels so as to characterize differences in neural processing based on distributed (rather than localized) activation patterns. This allows the detection of smaller effects, e.g. produced by perceptual differences between stimulus categories (Rasmussen and Williams, 2006; Tipping, 2001). The analyses involve a training stage in which the multivariate model is estimated based on a subset of the obtained data (training dataset) and a testing stage where the reliability of the model is assessed based on another subset of the obtained data (test dataset). The training stage is accomplished by a machine learning algorithm that aims at disclosing a

relationship between brain activation and experimental conditions. The testing phase is fundamental in assessing the validity of the model. Probabilistic models (such as Relevance Vector Machines (Formisano et al., 2008b)) and Gaussian Processes (Valente et al., 2011), are particularly suited for these applications, as they are designed to prevent over-fitting of the training data and have already proven considerably accurate in decoding brain states from fMRI measurements.

Specific Aims and outline of this thesis

Sound categories can be characterized by a unique mix of basic physical sound properties and higher-order harmonic information, or timbre. Interestingly, the auditory system is capable to categorize sounds that produce different timbres even when the lower level properties of these sounds are relatively similar (e.g. similar notes played on different instruments). Previous research supports a hierarchical model in which the processing of sound features relevant for sound recognition proceeds through a number of functionally specialized brain areas before culminating in category-selective processing modules that operate in the 'what' auditory cortical pathway.

The first two studies presented in this thesis challenge this hierarchical model and underline an alternative model that postulates rather parallel and distributed processing mechanisms in the human auditory system. The first experiment (Chapter 2) investigates whether neural representations of highly controlled natural sounds belonging to different categories can be differentiated by comparing BOLD responses to the different sounds using univariate, as well as multivariate methods.

The experiment described in Chapter 3 aims at investigating the temporal aspects of brain activation during natural sound perception. To that end MEG is used in combination with the stimuli from experiment 1. Compared to experiment 1, the physical differences between these stimuli are further minimized so as to create ambiguous stimuli that still evoke categorically different percepts.

As mentioned before, most natural auditory scenes contain many different, overlapping sounds that typically belong to different categories, e.g. male and female voices at a cocktail party. Fortunately, the auditory system may allow listeners to attend selectively to a single sound source and thereby enhance that sound's audibility. However, it is still unclear how to the auditory system achieves this feat. Besides timbre, another important cue for segregating a sound source from simultaneous sources is the spatial location of the source. The fMRI study in Chapter 4 investigates neural mechanisms for auditory stream segregation based on spatial cues, using binaural in-ear recordings of natural mixtures of voices and environmental sounds. In the fMRI study presented in Chapter 5, even more complex sound mixtures are employed to study the brain mechanisms underlying auditory scene analysis. Studio recordings from a band playing two pieces of music are used as stimuli during fMRI recordings. Specifically, the recordings are presented either separately (as individual instruments) or

together (i.e. as a composite mix) to investigate auditory stream segregation, during music perception. Using advanced data analysis methods (massively multivariate regression) we estimate the auditory cortical representations of a sound source (i.e. a musical instrument or a voice) that are robust to changes of the acoustic environment.

References

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., and Grady, C. L. (2001). "What" and "where" in the human auditory system. Proc Natl Acad Sci U S A *98*, 12301-12306.

Anderson, L. A., Christianson, G. B., and Linden, J. F. (2009). Stimulus-specific adaptation occurs in the auditory thalamus. J Neurosci *29*, 7359-7363.

Bizley, J. K., Walker, K. M., Silverman, B. W., King, A. J., and Schnupp, J. W. (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. J Neurosci *29*, 2064-2075.

Bregman, A. S. (1990). Auditory scene analysis (Cambridge, MA, MIT Press).

Cohen, D. (1972). Magnetoencephalography: detection of the brain's electrical activity with a superconducting magnetometer. Science *175*, 664-666.

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008a). "Who" is saying "what"? Brain-based decoding of human voice and speech. Science *322*, 970-973.

Formisano, E., De Martino, F., and Valente, G. (2008b). Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Magn Reson Imaging *26*, 921-934.

Formisano, E., Kim, D. S., Di Salle, F., van de Moortele, P. F., Ugurbil, K., and Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron *40*, 859-869.

Friston, K. J., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C., Frackowiak, R. S., and Turner, R. (1995). Analysis of fMRI time-series revisited. Neuroimage *2*, 45-53.

Griffiths, T. D. (2003). Functional imaging of pitch analysis. Ann N Y Acad Sci *999*, 40-49.

Hackett, T. A., Stepniewska, I., and Kaas, J. H. (1998). Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. J Comp Neurol *394*, 475-495.

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. (1993). Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. Reviews of Modern Physics *65*, 413.

Hari, R., Parkkonen, L., and Nangini, C. (2010). The brain in time: insights from neuromagnetic recordings. Ann N Y Acad Sci *1191*, 89-109.

Kaas, J. H., and Hackett, T. A. (1999). 'What' and 'where' processing in auditory cortex. Nat Neurosci *2*, 1045-1047.

Lee, C. C., Imaizumi, K., Schreiner, C. E., and Winer, J. A. (2004). Concurrent tonotopic processing streams in auditory cortex. Cereb Cortex *14*, 441-451.

Lewis, J. W., Brefczynski, J. A., Phinney, R. E., Janik, J. J., and DeYoe, E. A. (2005). Distinct cortical pathways for processing tool versus animal sounds. J Neurosci *25*, 5148-5158.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. Nature *412*, 150-157.

Ogawa, S., Menon, R. S., Tank, D. W., Kim, S. G., Merkle, H., Ellermann, J. M., and Ugurbil, K. (1993). Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging. A comparison of signal characteristics with a biophysical model. Biophys J *64*, 803-812.

Rasmussen, C. E., and Williams, C. K. I. (2006). Gaussian Processes for Machine Learning, The MIT Press).

Rauschecker, J. P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc Natl Acad Sci U S A *97*, 11800-11806.

Riecke, L., van Opstal, A. J., Goebel, R., and Formisano, E. (2007). Hearing illusory sounds in noise: sensory-perceptual transformations in primary auditory cortex. J Neurosci *27*, 12684-12689.

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., and Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat Neurosci *2*, 1131-1136.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., and Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. Curr Biol *19*, 498-502.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. J Mach Learn Res *1*, 211-244.

Valente, G., De Martino, F., Esposito, F., Goebel, R., and Formisano, E. (2011). Predicting subject-driven actions and sensory experience in a virtual world with relevance vector machine regression of fMRI data. Neuroimage *56*, 651-661.

CHAPTER 2

Sound categories are represented as distributed patterns in the human auditory cortex

Based on:

Staeren, N., Renvall, H., De Martino, F., Goebel, R., Formisano, E. Sound categories are represented as distributed patterns in the human auditory cortex. Current Biology (Volume 19, Issue 6, 498-502, 05 March 2009).

Summary

How does the brain recognize the sounds that populate our daily life? Previous research supports a hierarchical model of 'what' auditory cortical processing with category-selective modules. Processing of sound features relevant for sound recognition is assumed to proceed through a number of functionally-specialized areas, culminating in cortical modules where category-specific processing is carried out. Here we challenge this model by combining functional MRI and a novel machine learning algorithm, which is able to reveal both local as well as distributed neural representations. Sounds from four categories (cats, female singers, acoustic guitars, and tones) were controlled for their time-varying spectral characteristics and presented to subjects at three different pitch levels. Sound category information - not detectable using voxel-by-voxel analysis - could be detected and mapped with multivoxel pattern analyses. Processing of sound 'category' was spatially distributed over a large expanse of the supratemporal cortices, whereas a more localized pattern was observed for processing of 'pitch' laterally to primary auditory areas. Our findings indicate that distributed neuronal populations within the human auditory areas entail categorical representations of sounds, beyond their physical properties. A 'categorical' representation of a sound emerges from the joint encoding of information occurring not only in this small set of higher-level selective areas but also in the auditory areas conventionally associated with lower-level auditory processing.

Introduction

The ability to recognize sounds allows humans and animals to efficiently detect behaviorally relevant events, even in the absence of visual information. Anatomical and invasive electrophysiological studies in the macaque monkey (Kaas and Hackett, 1999; Rauschecker and Tian, 2000; Romanski et al., 1999) have suggested that auditory information relevant for sound recognition ("what") and localization ("where") is processed in two specialized and anatomically segregated streams of cortical areas. These processing streams originate in the anterior and posterior parts of the auditory cortex, respectively, and project to non-spatial and spatial domains of the frontal cortex. In humans, lesion (Adriani et al., 2003), electrophysiological (De Santis et al., 2007) and functional imaging studies (Alain et al., 2001; Arnott et al., 2004; Scott, 2005) have proposed the existence of similar streams for 'what' and 'where' auditory processing. Furthermore, specialized sub-systems for processing of other dimensions of auditory information, e.g. "how" (Belin and Zatorre, 2000) and "do" (Warren et al., 2005b), have been suggested.

The human auditory 'what' processing stream seems to include regions in the superior temporal cortex, located laterally to the primary auditory fields in the Heschl's gyrus (HG) (Formisano et al., 2003) and extending along the posterior-anterior direction of the superior temporal gyrus (STG) and sulcus (STS) (Alain et al., 2001; Warren and Griffiths, 2003). Processing of sound features relevant for sound recognition is assumed to proceed hierarchically through a number of functionally-specialized areas in this stream, culminating in cortical modules where category-specific processing is carried out. So far, strongest evidence for this modular model of functional architecture comes from fMRI studies that employed human and animal vocalizations as stimuli. Regions in the bilateral upper bank of the STS and adjacent STG exhibit a larger blood oxygenation level dependent (BOLD) response to vocal sounds than to non-vocal human-generated sounds (Belin et al., 2004; Belin et al., 2000; Warren et al., 2006). Similarly, the middle portions of the left and right STG (mSTG) are activated more during the categorization of animal vocalizations than tool sounds (Lewis et al., 2005). Recently, localized voice-selective BOLD responses have also been reported in the monkey cortex (Petkov et al., 2008). However, detailed functional architecture underlying the early stages of cortical processing of auditory 'what' information remains open. For example, it is not established whether these auditory regions are specialized for processing of human (and animal) vocalizations, or whether they account for a more general representation of sound categories, with voices being, for reasons both of acoustical complexity and behavioral relevance, the most prominent case. Results from studies using sounds other than voices have been less conclusive with respect to the early processing stages of the putative 'what' auditory stream. Indeed, previous studies that employed categorical comparisons between non-vocal sounds reported increased activation for these sounds in regions outside the areas that are typically defined as 'auditory'. For example, environmental sounds activated preferentially the bilateral posterior middle temporal gyrus (pMTG) (Lewis et al.,

2004) and hand-manipulated tool sounds a widespread, predominantly left-hemispheric network including frontal and parietal areas of the 'mirror-neuron system' (Lewis et al., 2005). These regions can be considered multimodal in terms of both anatomical and functional properties, and they probably represent a later processing stage than the supratemporal regions surrounding HG.

In the present high-resolution (2 x 2 x 2mm^3) fMRI study, we investigated the representation and processing of auditory categories within the human supratemporal cortex. In particular, we asked whether the areas around the primary auditory cortex would code for sound categories irrespective of their physical attributes, and if so, whether these representations would be localized in specialized areas or rather distributed across the auditory cortex.

Our investigation differs from previous studies of the 'what' auditory processing stream in terms of both stimulus design and data analysis strategy. First, sounds from different categories tend to differ also acoustically: Thus changes in the cortical responses between categories may also reflect merely their acoustic properties. Use of synthetic sounds would allow a more precise control over the acoustic properties of the stimuli (Patterson et al., 2002; Warren et al., 2005a). However, natural and synthetic sounds unavoidably differ in terms of ecological validity and familiarity, properties that are relevant for auditory neurons (Nelken, 2004; Wang et al., 2005). Ideally, one would like to compare cortical responses to sounds from different natural categories that are acoustically as similar as possible. Along these lines, we selected sounds from three 'real life' categories (female voices, cats, guitars) that were originally acoustically similar: All sounds were tonal with same fundamental frequency and similar harmonic structure (see Figure 1 and Methods). Besides being matched in terms of various physical properties like duration, root mean-square (RMS) power and temporal envelope, our stimuli were further manipulated by matching the temporal profile of their fundamental frequencies. This novel stimulus manipulation is particularly relevant as it ensured that the perceptual "pitch" dimension, mainly dependent on the sound fundamental frequency, was matched across categories.

Second, we employed an advanced analysis strategy based on an iterative machine learning algorithm (De Martino et al., 2008) that allows modeling of spatially distributed as well as localized response patterns. All previous studies on the 'what' auditory processing stream have utilized statistical univariate contrast-based analyses which are inherently bound to produce results in terms of 'specialization' or 'selectivity' for a certain stimulus attribute or category. Contrast-based methods can detect only localized surplus of hemodynamic activity for one condition compared with another, therefore ignoring the potential information of non-maximal responses. In an fMRI study of the object-vision pathway, Haxby and colleagues (Haxby et al., 2001) demonstrated that information on visual categories is not only encoded in the maximally responsive regions, but also in a spatially wide and distributed pattern of responses in the ventrotemporal cortex (the visual 'what' stream). Whether a

similar situation holds for the 'what' auditory processing stream is not known. For example, tool sounds that evoke smaller responses than voices in the superior temporal areas, may still exhibit response patterns that "code" for the category as informatively as the larger responses evoked by human or animal voices. Utilizing our recursive method for multivoxel pattern analysis we can directly address the issue of localized vs. distributed coding of auditory categories in STS/STG.

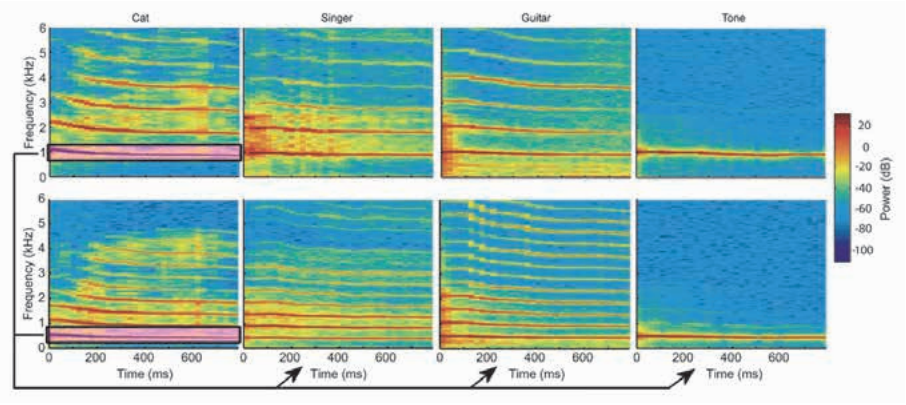

Figure 1. Spectrograms of exemplary stimuli. The four stimulus categories at High (920 Hz; top) and Medium (480 Hz; bottom) fundamental frequency levels. The time-varying fundamental frequency of the cat sound (purple rectangle) was imposed onto the other stimuli. The harmonic structure of the sounds was modified accordingly.

Results

During the fMRI measurements, subjects (n = 8) listened to sounds from three 'real life' categories (Singers, Cats, Guitars) and synthetic control sounds (Tones). All sounds were delivered binaurally via headphones in blocks of four at a comfortable listening level, using a clustered-volume acquisition technique that allowed for presentation of auditory stimuli in silence between subsequent volume acquisitions (see Experimental Procedures). Sounds within a block were from the same category and had the same of three possible fundamental frequencies (250 Hz = Low, 480 Hz = Middle and 920 Hz = High), resulting in altogether twelve experimental conditions. Examples of the stimuli can be found as Supplementary Audio files online.

Univariate statistical analysis

Figure 2 shows the responses to Singers, Guitars, Cats, and Tone stimuli compared with the baseline for a representative subject S2. All stimuli evoked significant BOLD responses in a large expanse of the auditory cortex, including bilateral HG, STG, and the upper bank of STS. With conventional univariate statistical contrasts, consistent differences were detected in

the superior temporal regions only for the Cats vs. Tones comparison (see Figure 3). At a rather lenient voxel-wise threshold of P = 0.01 (uncorrected), this contrast revealed significant differences in six out of the eight subjects. Any other univariate contrasts did not lead to statistically significant effects. Our control on the acoustic sound properties presumably reduced the voxel-by-voxel differences of BOLD responses evoked by the different sound categories.

Multivariate pattern recognition - Learning of sound 'category'

After this initial analysis, we used a statistical pattern recognition approach and tested the hypothesis that the overall spatial patterns of observed responses would convey information on the sound being presented. In each subject, we conducted six pair-wise classification experiments in which sound-evoked response patterns were labeled according to their category (Singers, Cats, Guitars, Tones), irrespective of their fundamental frequency. We examined whether our learning algorithm, after being trained with a subset of labeled brain responses (20 trials), would accurately classify the remaining unlabeled responses (10 trials, see Methods).

For all classifications, the recursive algorithm was able to learn the functional relation between the sounds and corresponding evoked spatial patterns and classify the unlabeled sound-evoked patterns significantly above chance level (0.5), with a mean classification correctness across subjects of 0.69 for Singers vs. Guitars ($P = 2.8401 \cdot 10^{-4}$, two-sided t test, n = 8), 0.69 for Singers vs. Cats ($P = 2.5552 \cdot 10^{-5}$), and 0.70 for Guitars vs. Cats ($P = 2.6351 \cdot 10^{-4}$) (Figure 4, left). The mean classification for Singers vs. Tones, correctness was 0.73 ($P = 4.7427 \cdot 10^{-7}$), 0.69 for Guitars vs. Tones ($P = 1.3517 \cdot 10^{-4}$), and 0.85 for Cats vs. Tones ($P = 3.53 \cdot 10^{-6}$) (Figure 5, left). These results suggest that spatially distributed patterns encoded information on sound category in the superior temporal regions.

Our method for the multivariate analysis of response patterns allows generating discriminative maps, i.e. maps of the locations that contribute most to the discrimination of conditions (see Methods). Figures 4 and Figure 5 depict the discriminative group maps of the classification between categories and between each category and control tones, respectively. It is important to note that for Cats vs. Tones the discriminative regions overlapped with the regions identified by the univariate contrast (see Figure 3).

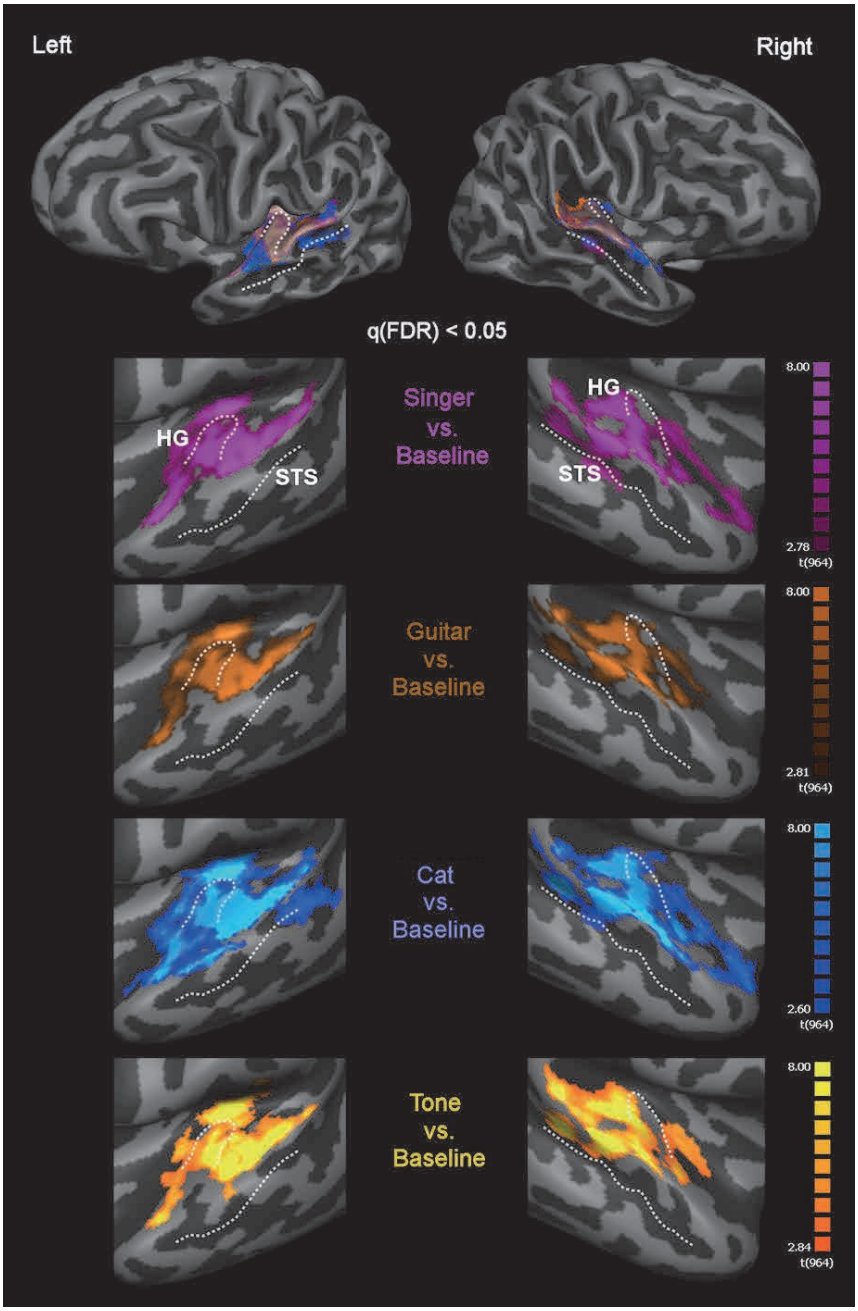

Figure 2. Auditory cortical responses to natural sounds (using univariate statistics). Activation maps for the contrasts between BOLD responses to Singer, Guitar, Cat, and Tone stimuli and the baseline in subject S5. All stimuli evoked significant BOLD responses (q(FDR) < 0.05) in a large expanse of the auditory temporal cortex, including the bilateral Heschl's gyrus (HG), the superior temporal gyrus (STG) and the superior temporal sulcus (STS).

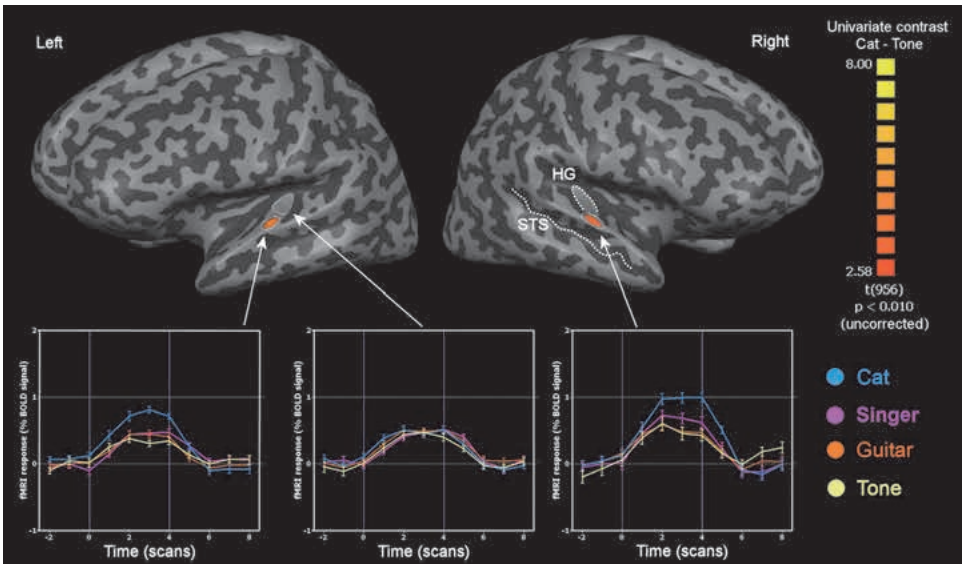

Figure 3. Univariate contrast Cats vs. Tones. Contrast map and the event-related averages illustrating the univariate statistical comparison of Cats vs. Tones. At a voxel-wise threshold of P = 0.01 (uncorrected), this contrast revealed significant differences in six out of the eight subjects (data in the Figure refer to subject S7). At the same threshold, all other univariate contrasts did not lead to statistically significant effects.

In order to quantify the consistency of the discriminative maps across subjects, group-level maps were generated (Figures 4, 5 and 6) by cortical realignment (Goebel et al., 2006) of individual discriminative maps. Single-subject maps included only voxels that "survived" the recursive elimination of irrelevant features in the algorithm (see Methods), and thus the group maps can be interpreted as a representation of spatial patterns that were consistently informative across subjects. A colored vertex indicates that the colored location was present in at least 60% (5/8) of the individual discriminative maps. At the group level, the distributed activation patterns that differentiated Singers from Guitars were located at the anterolateral HG, the planum temporale (PT), and the posterior STG and/or STS in the left hemisphere and at the lateral HG and the middle-posterior STG and/or STS in the right hemisphere. Singers were differentiated from Cats at the HS, the PT, and the posterior STG in the left hemisphere and at the middleposterior STG and the PT in the right hemisphere. Guitars were differentiated from Cats at the left anterolateral HG, the HS, and the posterior STG and at the right anterolateral HG, the PT, and the middle-posterior STG and/or STS. These results suggest that spatially distributed patterns encoded information on sound category in the superior temporal regions. The multivariate distributed activation patterns that discriminated between sound categories and tones are shown in Figure 5. Singers were differentiated from Tones in the left anterolateral HG, HS and posterior STG and in the right middle STG. Guitars were differentiated from Tones in the left middle-posterior STG, the right middle STG, and the right posterior STG/STS. Cats were differentiated from Tones in the left anterolateral HG,

HS, posterior STG/STS, and in the right-hemispheric anterolateral HG and medial posterior STG/STS. It is important to note that the regions for the Cats vs. Tones discrimination that achieved the highest classification correctness, overlapped with the regions identified by the univariate contrast (see Figure 3).

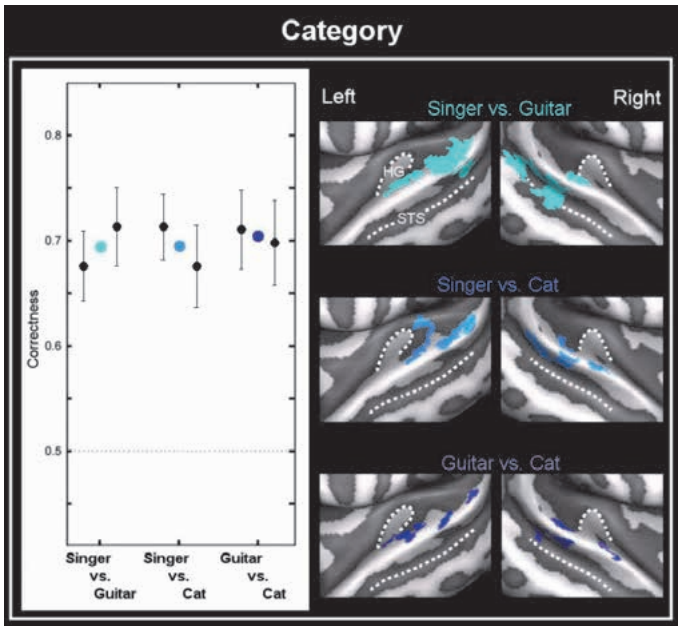

Figure 4. Multivariate pattern recognition - Learning of sound 'category'. Group averaged classification accuracies (left) and group discriminative maps (right) for between-category comparisons. For all binary discriminations, the black dots indicate the classification accuracy of test trials for each individual category, and the colored dots the classification accuracy averaged over the two categories. Error bars indicate the standard errors. For all classifications, the recursive algorithm was able to learn the functional relation between the sounds and corresponding evoked spatial patterns and classify the unlabeled sound-evoked patterns significantly above chance level (0.5). Discriminative patterns are visualized on the inflated representation of the auditory cortex resulting from the realignment of the cortices of the eight participants. A location was color-coded if it was present on the individual maps of at least five of the eight subjects.

Multivariate pattern recognition - Learning of sound 'fundamental frequency'

Because the stimuli were presented at three different fundamental frequency levels, we conducted a second analysis to investigate the regions that were most discriminative with respect to this second stimulus dimension. In this case, the same sound-evoked response patterns as used in the first analysis were labeled according to their fundamental frequency (High, Medium, Low), irrespective of their category. The recursive algorithm was then trained to discriminate the fundamental frequencies.

Figure 5 shows the resulting group discriminative maps and the corresponding correctness of 0.66 for Low vs. Medium ($P = 1.8187 \cdot 10^{-4}$, two-sided t test, n = 8), 0.68 for Low vs.

High ($P = 2.3 \cdot 10^{-3}$) and 0.68 for Medium vs. High ($P = 1.224 \cdot 10^{-4}$). As shown by the group discriminative maps, patterns related to fundamental frequencies were more clustered than the category discrimination maps, and they were circumscribed to the most lateral portion of HG. The group discriminative maps related to fundamental frequencies were more clustered than the category discriminative maps, and they were circumscribed to the most lateral portion of HG and/or HS bilaterally and to the posterior STG. This finding is in accordance with previous studies indicating this location as relevant for pitch processing using regular interval sounds (Griffiths, 2003; Patterson et al., 2002). Figure 7 summarizes the group discriminative maps obtained for the discrimination of categories (blue) and fundamental frequencies (red).

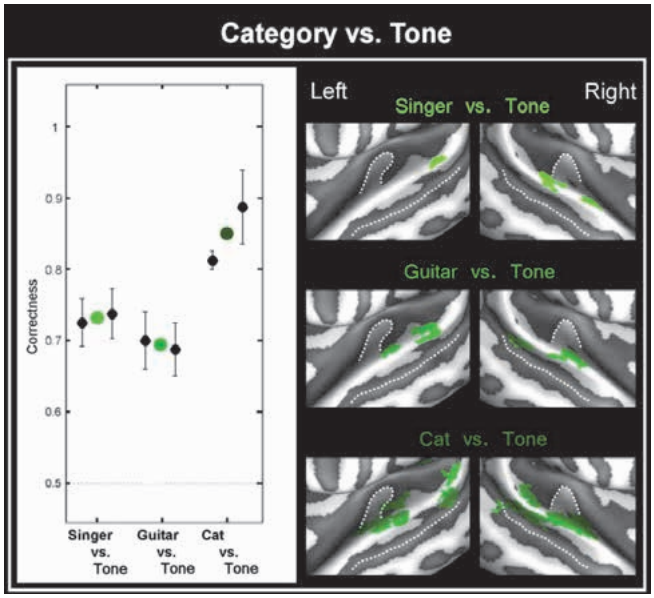

Figure 5. Multivariate pattern recognition – Classification of 'categories vs. tones'. Group averaged classification accuracies (left) and group discriminative maps (right) for the discrimination between categories (Singers, Guitars, Cats) and control Tones. For all binary discriminations, the black dots indicate the classification accuracy of test trials for each individual category, and the colored dots the classification accuracy averaged over the two categories. Error bars indicate the standard errors. For all classifications, the recursive algorithm was able to learn the functional relation between the sounds and corresponding evoked spatial patterns and classify the unlabeled sound-evoked patterns significantly above chance level (0.5). Discriminative patterns are visualized on the inflated representation of the auditory cortex resulting from the realignment of the cortices of the eight participants. A location was color-coded if it was present on the individual maps of at least five of the eight subjects.

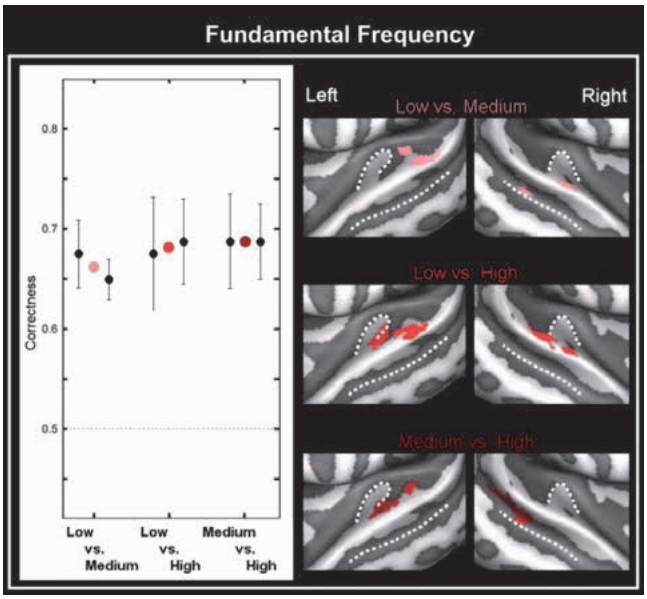

Figure 6. Multivariate pattern recognition - Learning of sound 'fundamental frequency'. Group averaged classification accuracies (left) and group discriminative maps (right) for between-frequency comparisons. For all binary discriminations, the black dots indicate the classification accuracy of test trials for each individual frequency, and the colored dots the classification accuracy averaged over the two frequencies. Error bars indicate the standard errors. For all classifications, the recursive algorithm was able to learn the functional relation between the sounds and corresponding evoked spatial patterns and classify the unlabeled sound-evoked patterns significantly above chance level (0.5). Discriminative patterns are visualized on the inflated representation of the auditory cortex resulting from the realignment of the cortices of the eight participants. A location was color-coded if it was present on the individual maps of at least five of the eight subjects.

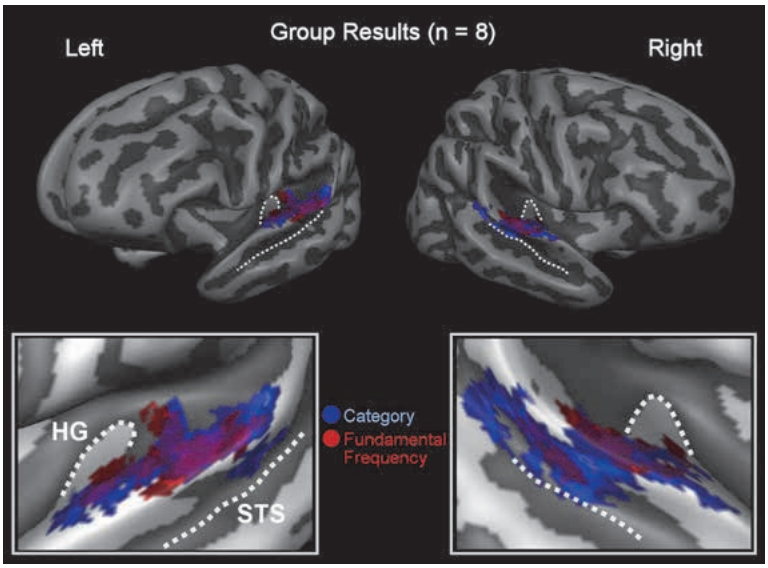

Figure 7. Comparison of discriminative maps. The cortex-based aligned group discriminative maps for category (blue) and fundamental frequency (red) discrimination. Category and fundamental frequency discriminative maps were obtained by the combination of the discriminative maps (logic OR) corresponding to the three binary classifications (Figures 4 and 6, respectively). A vertex was color-coded if it was present on the individual maps of at least five of the eight subjects. This corresponds to a false discovery rate-corrected threshold of $q = 7.9 \cdot 10^{-3}$ for the category map and $q = 2.6 \cdot 10^{-3}$ for the fundamental frequency map (see Methods). Note that the discrimination map for fundamental frequency was more clustered than that for category.

Discussion

Localized vs. distributed representation of sound categories

Our results indicate, similarly to the representation of visual object categories in the ventral temporal cortex (Haxby et al., 2001), that representations of sound categories in the superior temporal cortex are widely distributed and overlapping. The discriminative activation patterns extended bilaterally over a large expanse of the auditory cortex and included the anterior lateral portion of HG bilaterally, the posterior STG including the PT (mostly in the left hemisphere), the middle and anterior STG (mostly in the right hemisphere) and regions along the right STS. These locations overlap with – but are not limited to – locations that have been indicated in the previous investigations as functionally specialized areas for human (Belin et al., 2000) and animal (Lewis et al., 2005) vocalizations. In these studies, human voices were compared with other sound categories and phase scrambled sounds with similar global spectral aspects of the stimuli (Belin et al., 2000), and animal vocalizations were compared to tool sounds (Lewis et al., 2005). Thus the reported differences might reflect not only the real preference for a specific category but also unavoidable acoustic differences between test and control groups of stimuli.

In the present study, we have minimized the potential acoustical confounds. Our experimental sounds were controlled with respect to many acoustic dimensions, including their duration, average RMS level, amplitude envelope, harmonic-to-noise ratio (Boersma, 2001; Lewis et al., 2005) and the temporal profile of the sound spectrum. Removing most of the physical differences between categories diminished the differences between localized evoked BOLD responses, as reflected by the absence of between-category effects in our univariate analysis. Nevertheless, our iterative multivariate classification analysis showed that the activation patterns could be decoded into categories. Information in the spatially distributed patterns of activity may thus reflect a more abstract perceptual level of representation of sounds.

These findings put forward a revision of previous models of neuronal representation of complex sounds in the auditory cortex, which have implied a hierarchical functional architecture of auditory processing. In these models the superior temporal cortex is organized in specialized areas among which the neural processing of a sound hierarchically proceeds from the analysis of its low level physical constituents to higher perceptual dimensions. Within these models, auditory areas with a clear selectivity for a given category (e.g., voice) are seen as the functional units in which a more abstract representation of a sound is formed, independent of its specific acoustic features. However, it is a common observation in fMRI experiments that these 'higher level' areas show a vigorous BOLD response also to relatively simple stimuli (see the response to tones in Figure 2), implying sensitivity to low level properties of a sound as well. Based on our findings, we suggest that a 'categorical' representation of a sound emerges from the joint encoding of information occurring not only in this small set of higher-level selective areas but also in the auditory areas conventionally associated with "lower-level" auditory processing. This suggestion is not without prerequisites: The temporal auditory areas are anatomically heavily interconnected (Tardif and Clarke, 2001), and, even in the "early" auditory areas, neurons exhibit complex dependencies on the auditory input (Nelken, 2004; Wang et al., 2005). Furthermore, a distributed cortical coding of sound properties may explain why in human brain imaging several auditory regions have been implicated in the processing of many different auditory attributes (Griffiths and Warren, 2002). For example, PT has been attributed to motor transformation of auditory stimuli (Warren et al., 2005b), initial analysis of pitch (Patterson et al., 2002) and of auditory attributes relevant for sound localization and recognition (Griffiths and Warren, 2002).

Univariate vs. multivariate modeling of responses

Machine learning methods allow modeling of distributed patterns of cortical activations. These methods provide increased sensitivity compared with the conventional univariate statistical analysis by exploiting and integrating information from many spatial locations, thus allowing the detection of smaller effects, e.g. produced by perceptual differences between stimulus categories (Haynes and Rees, 2005; Kamitani and Tong, 2005).

We want to mention two aspects of our multivariate analysis. The first concerns the interpretation of accuracy levels, discriminative maps and their relation to univariate results. In cases in which significant differences between conditions could be detected already at single-voxel level, high classification accuracies were obtained. As expected, in these cases the multivariate discriminative maps and the univariate contrast maps overlapped (see, e.g., the Cats vs. Tones univariate contrast map in Figure 3, and the corresponding discrimination map in Figure 4). Discriminative maps, however, included additional sets of locations, whose joint activity and correlations were equally informative with respect to the classification of conditions. In the between-category discriminations, accuracy levels – albeit lower – were above chance in all our subjects and were obtained in the absence of significant univariate effects. Importantly, corresponding discriminative patterns were highly consistent across subjects. Taken together, these results suggest a genuine multivariate effect in which the accurate discrimination of categories was driven by information in spatially distributed patterns. Besides other methodological aspects (see below), the minimization of acoustical differences between categories and the absence of univariate effects may also explain why accuracy levels reached in our analyses are lower than those obtained in analogous analyses in the visual domain (Cox and Savoy, 2003; Haxby et al., 2001) in which physical differences between stimuli of visual categories were not accounted for.

Second, with our method, a multivariate analysis does not invariably lead to distributed results. For instance, in our analyses, re-labeling of the stimuli based on their fundamental frequency led the same learning algorithm used in the analysis of categories to find substantially different discriminative maps, with informative voxels clearly clustered in the lateral HG. In accordance with previous results (Griffiths, 2003), these findings support the notion that the processing of the fundamental frequency of a complex sound (and thus of perceptual 'pitch') is more localized. The discriminative maps of 'category' and 'fundamental frequency' overlapped substantially, thus suggesting that regions encoding relatively basic attributes of sounds, such as pitch, or higher level properties, such as category, are not mutually exclusive.

Limitations of present stimuli and extension to auditory scenes

The present stimuli were relatively simple and tonal by nature: For example, even though our Singers stimuli were real voices, their complexity was minimal compared with e.g. spoken language. Although this resulted in greater stimulus control, it also restricted the spectral richness and ecological validity of our stimuli. It remains to be proved that our findings are also valid for more complex natural sounds. It should be noted, however, that it will be challenging to carry out such an investigation while controlling for the acoustical differences of the sounds. To ensure enough acoustical variability to our stimuli, we presented all exemplars at three different fundamental frequencies. An accurate classification of novel sounds indicates that the machine-learning algorithm was able to extract a relation among stimuli (and corresponding activation patterns), which we assume to be at the level of 'category'. It

should be noted, however, that despite our efforts in equalizing low-level acoustic properties, the degree of acoustical similarities between sounds of the same category is higher than for sounds of different categories. It is thus possible that the level of representation driving the learning process may reflect the decoding of complex combination of spectral and temporal features that characterize what we have defined as 'sound category'. The question of high order representation of a natural sound may be addressed by testing the ability of a brain-based classifier to generalize its performance in realistic situations that require abstraction from low-level features, e.g. in recognizing a voice in a noisy scene after training the classifier with voices presented in silence.

Experimental Procedures

Subjects

We studied, with informed consent, one Belgian and eight Dutch subjects (mean age ± SD 24 ± 5 yrs; 8 females and one male; all right-handed). The subjects were undergraduate university students who were paid for their participation. Subjects had no history of hearing or neurological impairments, and were naïve to the experimental setup. The study received a prior approval by the Ethical Committee of the Faculty of Psychology and Neuroscience, University of Maastricht.

Auditory stimuli

The stimuli were 800-ms sounds (sampled at 44.1 kHz) from four sound categories: cats, singers (singing female voices), acoustic guitars and tones. Each category except the tones consisted of three different representatives (e.g. three different singers). All sounds were transposed to three different fundamental frequencies (250, 480 and 920 Hz), thus resulting in altogether twelve conditions. The values of fundamental frequencies were chosen so as to ensure that stimuli were clearly recognizable and to avoid pure octave pitch differences (e.g. 250, 500 and 1000 Hz).

To equalize the spectrotemporal profiles and the perceptual pitch of the stimuli, the time-varying fundamental frequency of the cat sounds was extracted on 25 time points within each stimuli with Praat software (Boersma, 2001) and applied continuously to all other sounds with Adobe Audition™. Note that not only the fundamental frequency of manipulated sounds was adjusted, but all related harmonics (see Figure 1 and online Supplementary Audio Files). Cat sounds were chosen as reference stimuli because of relatively small temporal variations in their fundamental frequency. The acoustic guitar and female singers were chosen as the other categories because for these sounds continuous pitch changes are natural (e.g., sliding in between two tones when singing, or bending a guitar string) and thus they were still clearly recognizable after the pitch matching procedure. Tones were used as control sounds. The sounds were low-pass filtered at 14 kHz for five subjects, and to further minimize the acoustical differences between sound categories, at 7 kHz for three subjects.

No significant differences between the results of these groups were found in the univariate and multivariate statistical analysis, and thus subjects were grouped together in reported results. The sound amplitude envelopes and average root-meansquare levels were matched using MATLAB 7.0.1 (The MathWorks, Inc., Natick, MA, USA). The harmonic-to-noise ratio (Boersma, 2001; Lewis et al., 2005) was significantly different only between tones and sound categories ($P < 0.001$), not between categories ($P > 0.05$).

Before the fMRI measurements, all subjects underwent a training session. Subjects were asked to listen to the stimuli until they subjectively felt they were able to clearly categorize the stimuli. Typically the subjects listened to all the sounds 2~3 times. Data from one subject were discarded from further analysis on the basis of incorrect interpretation of the task instructions. Hearing thresholds for different categories and pitch levels were tested individually for each subject, and stimuli were adjusted accordingly. Following the fMRI sessions (see below), subjects were enquired on the difficulty of attributing the stimuli to a given category during the scanning. All subjects indicated that categorization was easy for all stimuli.

fMRI measurements

Brain imaging was performed with a 3 Tesla Siemens Allegra (head setup) at the Maastricht Brain Imaging Center. In each subject, two runs of 488 volumes were acquired with a T2-weighted gradient-echo planar imaging (EPI) sequence (TR = 3610 ms, voxel size = 2 x 2 x 2 mm^3, TE = 30 ms, FOV 256 x 256; matrix size 128 x 128, 23 slices covering the perisylvian cortex). Each run consisted of 15 blocks per sound category and lasted approximately 30 min. Anatomical images were obtained using a 1 x 1 x 1 mm^3 resolution T1-weighted sequence between the functional runs.

During the measurements, the stimuli were delivered binaurally via MR compatible headphones (Commander XG, Resonance Technology, Northridge, CA) in blocks of four at a comfortable listening level. To minimize the effect of scanner noise, the sounds were presented during 1600-ms silent periods between 2000-ms scans; the 800-ms sounds were preceded and followed by a 400-ms silence, using a clustered volume EPI technique that allowed for presentation of auditory stimuli in silence between subsequent volume acquisitions (Jancke et al., 2002; Riecke et al., 2007; van Atteveldt et al., 2004). The stimuli within a block were from the same category and frequency level, resulting in altogether twelve experimental conditions. The experimental blocks had duration of 14.4 s. The conditions were repeated in a pseudo-random order, and were followed by rest period of identical length, at the beginning of which the subjects were asked to respond with a button press whether the last two sounds in the block were the same (50% of the catch trials). The response hand was alternated across subjects.

fMRI Data Analysis: pre-processing and univariate statistics

Functional and anatomical images were first analyzed with BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). Preprocessing consisted of slice scan-time correction

(using sinc interpolation), linear trend removal, temporal high-pass filtering to remove non-linear drifts of seven or less cycles per time course, and 3-dimensional motion correction. Temporal low pass filtering was performed using a Gaussian kernel with FWHM of two data points. Moderate spatial smoothing with a Gaussian kernel of FWHM of three millimeters was performed on the volume time series. Functional slices were co-registered to the anatomical data, and both data were normalized to Talairach space (Talairach and Tournoux, 1988).

Conventional univariate statistical analysis of the fMRI data was based on the general linear modeling (GLM) of the time series. For each subject, a design matrix was formed using a predictor for each stimulus category. The predicted time courses were adjusted for the hemodynamic response delay by convolution with a canonical double gamma) hemodynamic response function. Contrast maps were thresholded on the basis of False Discovery Rate (q = 0.05) when comparing sound categories with the baseline (Figure 2), or at an exploratory threshold of P = 0.01 (uncorrected for multiple comparison) in the case of direct comparison between sound categories (Figure 3).

fMRI Data Analysis: multivariate pattern recognition

Multivoxel patterns of sound-evoked BOLD responses were analyzed using a method that combines machine learning with an iterative, multivariate voxel selection algorithm, Recursive Feature Elimination (RFE) (De Martino et al., 2008). This method allows estimating maximally discriminative response patterns without a priori definition of regions of interest. In brief, starting from the entire set of measured voxels our method uses a training algorithm (least square support vector machine, ls-SVM) iteratively to eliminate irrelevant voxels and to estimate the informative spatial patterns. Correct classification of the test data increases, while features/voxels are pruned on the basis of their discrimination ability. We have recently validated and compared this method to other approaches of multivoxel pattern analysis and demonstrated its greater sensitivity by means of simulations. A short description of the method is given below, together with steps and parameters specific to the analysis of present data. A more complete account of the implementation and validation of the method can be found in (De Martino et al., 2008). Pre-processed functional time series were first divided into “trials” (one trial per block) and labeled either according to the category (learning of ‘category’) or the fundamental frequency (learning of ‘fundamental frequency’) of the sounds presented in the block. This gave rise, in each subject, to a total of 30 trials per condition for category discrimination, and 40 trials per condition for fundamental frequency discrimination. For each trial, a multivoxel pattern response was generated. An estimate of the response at every voxel was obtained by fitting a general linear model with one predictor coding for the trial response and one linear predictor accounting for a within-trial linear trend. The trial response predictor was obtained by convolution of a boxcar with a double gamma hemodynamic response function. The corresponding regressor coefficient (beta) was taken to represent the voxel trial response and responses from all voxels were combined to

form multivoxel patterns. Multivoxel pattern responses were analyzed using the iterative ls-SVM-based classification algorithm. For each pair of categories (or fundamental frequencies), trials were divided into a training set (20 trials per condition for the category discrimination and 30 trials per condition for the fundamental frequency discrimination) and a test set (10 trials per condition). The training set was used for estimating the maximally discriminative patterns with the iterative algorithm; the test set was only used to assess the correctness of classification of unseen trials (i.e. not used in the training).

Starting from all the cortical voxels included in a subject-by-subject defined anatomical mask (including temporal pole, STG, STS, MTG), the most active voxels per condition (as defined on the training set alone) were initially selected. The threshold for this initial activation-based voxel selection was optimized for each subject by using a cross validation within the training data, and the threshold ranged between 1000 and 1500 voxels per condition.

Voxels were further reduced using the iterative RFE algorithm. At each iteration, RFE included two steps. First, a subset of the training data (10 trials per condition for the category discrimination and 20 trials per condition for the fundamental frequency discrimination) was used to train an ls-SVM classifier. As a result of this training, a map coding for the relative contribution of each voxel to the discrimination of conditions (discriminative maps) was obtained as in (Mourao-Miranda et al., 2005). Second, these discrimination weights were ranked and voxels corresponding to the smallest ranking were discarded. Voxels with the highest discriminative values were used for training in the next iteration. These two steps were repeated ten times (Nit = 10, on different subsets of the training data), each time with a 30% reduction in the number of voxels. The correctness of the classification corresponding to the current set of voxels and the discriminative weights were assessed using the external test trials. The entire iterative procedure was repeated with cross validation ten times (Nsplits = 10), each time leaving out a different subset of trials per condition. The reported correctness for each single class and each binary comparison was computed as an average across the ten splits (Figure 4, 5 and 6). Single-subject discriminative maps corresponded to the voxel-selection level that gave the highest average correctness. These maps were then sampled on the reconstructed cortex of each individual subject and binarized in order to visualize only the best 20% of the vertices.

To examine the spatial consistency of the discriminative patterns across subjects, group-level discriminative maps were generated after cortex-based alignment (Goebel et al., 2006) of single-subject discriminative (binarized) maps (Formisano et al., 2008). In these group-level discriminative maps, a cortical location (vertex) was color-coded if it was present in the corresponding individual discriminative map of at least five of the eight subjects. Assuming that the discriminative maps for category and fundamental frequency follow a binomial distribution, the likelihood of finding the same locations by chance in five subjects corresponds to an "uncorrected" $p = 8.4 \cdot 10^{-4}$ for the category map and an "uncorrected" $p = 1.3 \cdot 10^{-4}$ for the fundamental frequency map. To account for the multiple tests performed to create

these maps, we calculated the proportion of expected false positive in each of the maps (False Discovery Rate, q) that correspond to these p values. This resulted in $q = 7.9 \cdot 10^{-3}$ for category and $q = 2.6 \cdot 10^{-3}$ for fundamental frequency. These q-values were computed using a statistical method that ensures robust estimates also in the case of discrete distribution of p-values and onesided tests (Pounds and Cheng, 2006).

Acknowledgements

We would like to thank Lars Riecke for his useful comments on the experimental design and on the manuscript. Funding for the present research was contributed to E. Formisano (Vernieuwingsimpuls VIDI) from the Netherlands' Organization for Scientific Research (NWO) and to H. Renvall from the Academy of Finland.

References

Adriani, M., Maeder, P., Meuli, R., Thiran, A. B., Frischknecht, R., Villemure, J. G., Mayer, J., Annoni, J. M., Bogousslavsky, J., Fornari, E., *et al.* (2003). Sound recognition and localization in man: specialized cortical networks and effects of acute circumscribed lesions. Exp Brain Res *153*, 591-604.

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., and Grady, C. L. (2001). "What" and "where" in the human auditory system. Proc Natl Acad Sci U S A *98*, 12301-12306.

Arnott, S. R., Binns, M. A., Grady, C. L., and Alain, C. (2004). Assessing the auditory dual-pathway model in humans. Neuroimage *22*, 401-408.

Belin, P., Fecteau, S., and Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. Trends Cogn Sci *8*, 129-135.

Belin, P., and Zatorre, R. J. (2000). 'What', 'where' and 'how' in auditory cortex. Nat Neurosci *3*, 965-966.

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

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International *5*, 341-345.

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage *19*, 261-270.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. Neuroimage *43*, 44-58.

De Santis, L., Clarke, S., and Murray, M. M. (2007). Automatic and intrinsic auditory "what" and "where" processing in humans revealed by electrical neuroimaging. Cereb Cortex *17*, 9-17.

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. Science *322*, 970-973.

Formisano, E., Kim, D. S., Di Salle, F., van de Moortele, P. F., Ugurbil, K., and Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron *40*, 859-869.

Goebel, R., Esposito, F., and Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Hum Brain Mapp *27*, 392-401.

Griffiths, T. D. (2003). Functional imaging of pitch analysis. Ann N Y Acad Sci *999*, 40-49.

Griffiths, T. D., and Warren, J. D. (2002). The planum temporale as a computational hub. Trends Neurosci *25*, 348-353.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science *293*, 2425-2430.

Haynes, J. D., and Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci *8*, 686-691.

Jancke, L., Wustenberg, T., Scheich, H., and Heinze, H. J. (2002). Phonetic perception and the temporal cortex. Neuroimage *15*, 733-746.

Kaas, J. H., and Hackett, T. A. (1999). 'What' and 'where' processing in auditory cortex. Nat Neurosci *2*, 1045-1047.

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. Nat Neurosci *8*, 679-685.

Lewis, J. W., Brefczynski, J. A., Phinney, R. E., Janik, J. J., and DeYoe, E. A. (2005). Distinct cortical pathways for processing tool versus animal sounds. J Neurosci *25*, 5148-5158.

Lewis, J. W., Wightman, F. L., Brefczynski, J. A., Phinney, R. E., Binder, J. R., and DeYoe, E. A. (2004). Human brain regions involved in recognizing environmental sounds. Cereb Cortex *14*, 1008-1021.

Mourao-Miranda, J., Bokde, A. L., Born, C., Hampel, H., and Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. Neuroimage *28*, 980-995.

Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. Curr Opin Neurobiol *14*, 474-480.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., and Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. Neuron *36*, 767-776.

Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., and Logothetis, N. K. (2008). A voice region in the monkey brain. Nat Neurosci *11*, 367-374.

Pounds, S., and Cheng, C. (2006). Robust estimation of the false discovery rate. Bioinformatics *22*, 1979-1987.

Rauschecker, J. P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc Natl Acad Sci U S A *97*, 11800-11806.

Riecke, L., van Opstal, A. J., Goebel, R., and Formisano, E. (2007). Hearing illusory sounds in noise: sensory-perceptual transformations in primary auditory cortex. J Neurosci *27*, 12684-12689.

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., and Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat Neurosci *2*, 1131-1136.

Scott, S. K. (2005). Auditory processing--speech, space and auditory objects. Curr Opin Neurobiol *15*, 197-201.

Talairach, J., and Tournoux, P. (1988). Co-Planar Stereotactic Atlas of the Human Brain (Stuttgart, Thieme).

Tardif, E., and Clarke, S. (2001). Intrinsic connectivity of human auditory areas: a tracing study with DiI. Eur J Neurosci *13*, 1045-1050.

van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (2004). Integration of letters and speech sounds in the human brain. Neuron *43*, 271-282.

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

Warren, J. D., and Griffiths, T. D. (2003). Distinct mechanisms for processing spatial sequences and pitch sequences in the human auditory brain. J Neurosci *23*, 5799-5804.

Warren, J. D., Jennings, A. R., and Griffiths, T. D. (2005a). Analysis of the spectral envelope of sounds by the human brain. Neuroimage *24*, 1052-1057.

Warren, J. D., Scott, S. K., Price, C. J., and Griffiths, T. D. (2006). Human brain mechanisms for the early analysis of voices. Neuroimage *31*, 1389-1397.

Warren, J. E., Wise, R. J., and Warren, J. D. (2005b). Sounds do-able: auditory-motor transformations and the posterior temporal plane. Trends Neurosci *28*, 636-643.

CHAPTER 3

Of cats and women:

Temporal dynamics in the right temporoparietal cortex reflect auditory categorical processing of vocalizations

Corresponding publication:
Renvall, H., Staeren, N., Siep, N., Esposito, F., Jensen, O., Formisano, E. Of cats and women: Temporal dynamics in the right temporoparietal cortex reflect auditory categorical processing of vocalizations. Neuroimage (Volume 62, Issue 3, 1877-1883, 19 June 2012)

Summary

Understanding the temporal dynamics underlying cortical processing of auditory categories is complicated by difficulties in equating temporal and spectral features across stimulus classes. In the present magnetoencephalography (MEG) study, female voices and cat sounds were filtered so as to match in most of their acoustic properties, and the respective auditory evoked responses were investigated with a paradigm that allowed us to examine auditory cortical processing of two natural sound categories beyond the physical make-up of the stimuli. Three cat or human voice sounds were first presented to establish a categorical context. Subsequently, a probe sound that was congruent, incongruent, or ambiguous to this context, was presented. As an index of a categorical mismatch, MEG responses to incongruent sounds were stronger than the responses to congruent sounds at ~250 ms in the right temporoparietal cortex, regardless of the sound category. Furthermore, probe sounds that could not be unambiguously attributed to any of the two categories ("cat" or "voice") evoked stronger responses after the voice than cat context at 200–250 ms, suggesting a stronger contextual effect for human voices.

Our results suggest that categorical templates for human and animal vocalizations are established at ~250 ms in the right temporoparietal cortex, likely reflecting continuous online analysis of spectral stimulus features during auditory categorizing task.

Introduction

The ability to rapidly recognize and categorize sounds is essential, not only for understanding and reacting to our surroundings, but for daily communication and social interaction. Studies in macaque monkeys have suggested that auditory information relevant for sound recognition in general is processed in a specialized and anatomically segregated stream of cortical areas (Kaas and Hackett, 1999; Rauschecker and Tian, 2000; Romanski et al., 1999). Correspondingly in humans, sound recognition activates regions located laterally to the Heschl's gyrus and extending along the posterior–anterior direction of the superior temporal gyrus (STG) and sulcus (STS) (Alain et al., 2001; Warren and Griffiths, 2003). Within these areas, sound categories are encoded in a spatially distributed manner (Formisano et al., 2008; Staeren et al., 2009).

In humans, both animal and human vocalizations constitute rapidly and effortlessly recognizable auditory categories that are learned early in childhood and share many spectorotemporal features. Vocalizations activate specific auditory networks: Regions in the bilateral STS and STG exhibit a larger blood-oxygenation-level-dependent response to vocal than to non-vocal human sounds (Belin et al., 2004; Belin et al., 2000; Warren et al., 2006), and the middle portions of the STG are bilaterally more activated during the categorization of animal vocalizations than tool sounds (Lewis et al., 2005). Furthermore, sub-regions at these areas show species-specific reactivity to vocalizations (Fecteau et al., 2004).

In functional magnetic resonance imaging (fMRI) studies, minimizing the low-level acoustic differences between stimuli abolishes conventional univariate differences between responses to different sound categories (Staeren et al., 2009). Exemplars of separate categories differ from each other temporospectrally, and time-sensitive electroencephalographic (EEG) and magnetoencephalographic (MEG) responses are especially sensitive to such deviations. In a recent EEG study, responses to human voices differed from those to bird songs and environmental sounds at ~200 ms bilaterally at the fronto-temporal electrodes, but the results were speculated to be at least partly due to differences between the experimental stimuli (Charest et al., 2009). Another EEG study, in which the sound spectrograms and power spectra did not statistically significantly differ between sound categories, demonstrated stronger activity to human than animal vocalizations at 169–219 ms over the right temporal areas (De Lucia et al., 2010). However, the same ~200-ms time window has been related to general processing of spectral fine structure of any complex sound (Altmann et al., 2008), and the nature of auditory categorical processing has remained unclear.

Here we used MEG in combination with acoustically well-controlled human and cat vocalizations to study cortical processing of auditory categories beyond the processing of low-level features. As an important addition to previous studies, the temporal profiles of our stimuli were equated for their harmonic structures. This manipulation ensures that the sounds have a similar "perceptual pitch" profile over time, behaviourally relevant for sound categorization (Staeren et al., 2009). Furthermore, we used an adaptation paradigm in which

exact same stimuli could be presented in different contexts. Based on a predictive coding account of auditory adaptation (Friston, 2005; Garrido et al., 2008; Jaaskelainen et al., 2004; Wacongne et al., 2011), we hypothesized that sounds incongruent to the preceding context, would produce - in the superior temporal cortex - stronger responses than congruent sounds as a marker of a categorical mismatch. Finally, we probe and compare these categorical adaptation effects for the two different contexts ("voice" and "cat") with acoustically identical target sounds that could not be unambiguously attributed to any of the two categories.

Materials and methods

Subjects

We studied, with informed consent, 8 adults (mean ± SEM age 28 ± 1 yrs; 3 females, 5 males; 7 right-handed and one ambidextrous). None of the subjects had a history of hearing or neurological impairments, and the study received a prior approval by the Ethical Committee of the Faculty of Psychology and Neuroscience, Maastricht University.

Auditory Stimuli and Experimental Design

One cat (meowing) and one voice sound (singing female) were selected from the stimulus set used in Staeren et al. (2009), on the basis of their close resemblance in harmonics-to-noise ratios (Boersma, 1993; Lewis et al., 2005; Murray et al., 2006) and power spectra. To further minimize the spectrotemporal differences between the stimuli, the time-varying fundamental pitch of the cat sound was extracted at 25 time points (in ~30 ms steps) within the stimulus with Praat software (Boersma, 2001) and applied to the voice sound using Adobe Audition™. Sounds were then low-pass (LP) filtered at 13 cutoff frequencies; the LP frequencies varied in steps of 100 Hz between 500 and 900 Hz, and in steps of 200 Hz between 900 and 2500 Hz. To add more variation to the stimuli, they were transposed to five different fundamental frequencies between 230–260 Hz. These procedures resulted in 65 stimuli for each of the two categories (5 pitch levels × 13 frequency ranges). The stimuli lasted for 780 ms, and they were equalized for their mean intensities with MATLAB 7.0.1 ™ (The MathWorks, Inc., Natick, MA, USA). Differences in stimulus amplitude envelopes between cat and voice stimuli were minimized by using 10-ms moving-average windows, to an extent not to disturb original sound quality. The remaining amplitude differences were tested by analyzing the sound intensities in 20-ms steps at 0–220 ms from the beginning of the stimuli: the stimulus intensities did not differ statistically significantly between the cat and voice stimuli ($P > 0.09$).

The stimuli were tested behaviourally in 14 subjects who did not participate in the final experiment. In these behavioural tests, subjects were first familiarized with six easily recognizable representatives from both categories together with visual information about the sound category (Presentation 9.3™, Neurobehavioral Systems, Inc., Albany, CA, USA). Then, they were instructed to carefully listen to the sounds presented at 2 s interstimulus intervals

(ISI), and to decide whether the sound was a voice or a cat stimulus. Subjects were asked to be as accurate and fast as possible, and their ratings were reported through button presses. After a few practise trials, each stimulus (65 per category) was presented nine times.

At the largest bandwidths, the stimuli sounded very natural and, correspondingly, they were easily recognized as representatives of their category, while narrowing the bandwidth gradually affected the behavioral response. On the basis of the results, nine cat/voice stimulus pairs with similar recognition accuracies and reaction times between categories were selected as "easy". These sounds consisted of LP levels 1500 Hz (at two different pitch levels), 1900 Hz (three pitch levels), and 2300 Hz (four pitch levels). In addition, the voice sounds that were LP-filtered at 500 Hz (four pitch levels) resulted in behavioural responses at chance level, and they were selected as "ambiguous". Examples of the stimuli and their spectrograms are presented in Figure 1. Despite the efforts to minimize the spectrotemporal differences between stimulus categories, the easily recognizable female voice stimuli contained more energy at ~1000–1500 Hz than the cat vocalizations throughout the stimulus duration (see Fig. 1a and 1b). Although the ambiguous stimuli were modified from the voice stimuli by LP filtering at 500 Hz and thus their resembled more closely the voice stimuli in their amplitude behavior, their spectrotemporal structure was rather flat at 0–500 Hz and did not contain the upper harmonics that were characteristics for both the easy voice and cat stimuli.

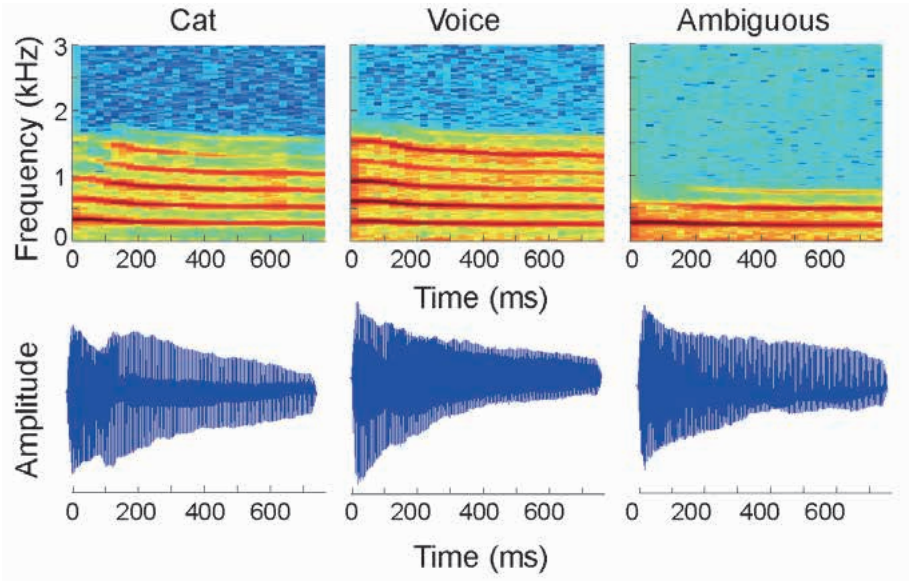

Figure 1. Spectrograms of exemplary cat and voice stimuli (both low-pass filtered at 1900 Hz), and of ambiguous stimuli (voice sound, low-pass filtered at 500 Hz). The time-varying fundamental frequency of the cat sound was extracted and imposed onto the voice stimuli; All the harmonics of the voice sounds were modified accordingly.

During the MEG session, the behavioural responses were too scarce for statistical inference. Therefore, in a separate behavioural session prior to the MEG experiment, all subjects underwent a short behavioural session (Presentation 9.3™). First the subject listened twice to all nine "easy" cat and voice stimuli presented with an ISI of 2 s, together with visual information on the stimulus category. Subsequently, the same stimuli were presented randomly three times without visual aid and interspersed with the ambiguous stimuli, and the subject was asked to respond with a button press whether the stimulus was a cat or a female voice. Finally, the subjects listened to the sounds as they would be presented in the MEG experiment, i.e. four sounds in a row, and they were asked to respond after each trial whether the all four sounds belonged to the same category (yes/no).

The percentage of correct cat and voice sound recognition was $\geq 97 \pm 2$ % (mean ± SEM). Subjects' responses to the ambiguous sounds were at the chance level: The percent correct (the subject responded 'voice') was 39 ± 12% when the sounds were presented after cat sounds, and 63 ± 14% after voice sounds ($p > 0.35$ compared with 50%), and the responses did not differ statistically significantly from each other ($p = 0.15$).

MEG experiment

In the MEG experiment, the sounds were delivered to the subjects binaurally at a comfortable listening level through plastic tubes and ear pieces. They were presented in trains of four, and the subject's task was to attend to all sounds carefully, and decide whether the sounds belonged to the same category (cat or voice). The experiment is described schematically in Figure 2. The stimuli within a train were presented with ISIs of 600 ms (from offset to onset), resulting in a trial duration of 4920 ms, and they were followed by an inter-trial interval of 2700 ms.

The experiment consisted of six conditions utilizing the stimuli described above (nine voice sounds, nine cat vocalizations and four ambiguous sounds). In the congruent conditions, four cat (or voice) sounds were presented in a row. In the incongruent conditions, three voice (cat) sounds were followed by a cat (voice) sound. In the ambiguous conditions, three voice (or cat) stimuli were followed by an ambiguous stimulus. To minimize build-up of purely acoustic memory traces during the trials and to avoid mismatch responses elicited by infrequent sounds among otherwise monotonous stimulation (Näätänen, 1992), the three first stimuli in a train were selected each from a different filtering level. The last sound in a row could be either from the same or different filtering level as the preceding third sound; MEG responses were pooled across the different filtering and pitch levels. The different stimulus trains were presented in a random order, and the same condition was not allowed to occur more than twice in succession.

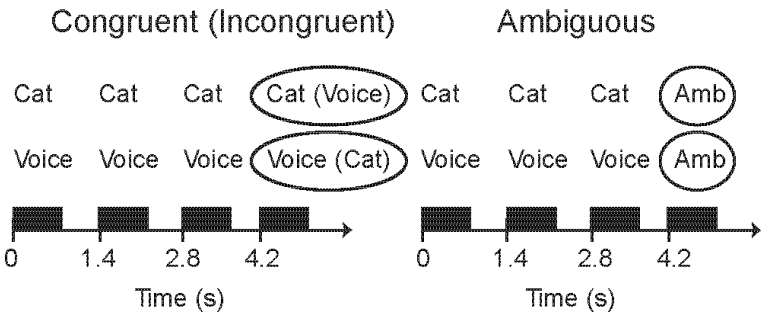

Figure 2. Schematic presentation of the incongruent and ambiguous experimental trials. Note that the stimuli within a trial varied both in their LP filtering and pitch levels (see text).

In 7% of the trials, a question mark appeared 1 s after the last stimulus, and the subject was required to respond by lifting her/his index or middle finger whether the sounds belonged to a same category (yes/no). The subsequent trials were discarded from the analysis. The response hand was alternated across subjects, and to minimize possible motor contamination on the data, subjects were instructed to keep their hand relaxed during the experiment. To prevent subjects' deciding on the last stimulus only, 7% of the trials were "catch trials" in which the incongruent stimulus occurred at the first, second or third stimulus position. These responses were also removed from the data analysis.

Auditory evoked fields were recorded in a magnetically shielded room using a whole-head MEG system (VSM/CTF Systems Inc., Port Coquitlam, Canada) with 275 axial gradiometers. Three head-position-indicator coils were attached at anatomical landmarks (the left and right ear canals and the nasion). The head position with respect to the sensor array was determined by feeding current to the marker coils and measuring their positions with respect to the sensory array before and after the measurements.

The MEG signals were low-pass filtered at 300 Hz and digitized at 1200 Hz, and averaged offline with two time scales: i) from 200 ms before the onset of the whole stimulus block to 1000 ms after the onset of the last (4th) stimulus, and ii) from 200 ms before the onset of each stimulus to 1000 ms after it. The averaged signals were digitally low-pass filtered at 40 Hz, and a prestimulus baseline of 200 ms was applied.

The experiment was conducted in 5 blocks, each lasting ~10 min. During the experiment each of the six conditions (two congruent, two incongruent and two ambiguous conditions) was repeated 70 times. Horizontal and vertical electro-oculograms were recorded to discard data contaminated by eye blinks and movements; ~60–70 artifact-free responses were averaged per condition.

MEG sensor-level signals

For an initial estimate of the experimental effects, the responses to whole stimulus blocks were first analysed at the sensor level. To simplify the analysis, a planar gradient was esti-

mated for each channel from the neighbouring channels (Medendorp et al., 2007); Planar gradients give the maximum signal just above the source area (Hämäläinen et al., 1993). Root mean square of the horizontal and vertical planar gradient fields was then calculated (combined planar gradient). Subsequently areal mean averages were calculated over the central, left and right temporal, left and right frontal, and left and right occipito-parietal regions.

Source analysis: equivalent current dipole modeling

For source analysis, the head was modelled as a homogeneous spherical volume conductor. The model parameters were optimised for the intracranial space obtained from MR images that were available for all subjects. The neurophysiological responses were analyzed by first segregating the recorded sensor-level signals into spatiotemporal components, by means of manually-guided multi-dipole current modelling (equivalent current dipole, ECD; (Hämäläinen et al., 1993). The analysis was conducted separately for each subject using Elekta Neuromag (Elekta Oy) software package, following standard procedures (Hansen et al., 2010; Salmelin et al., 1994). The parameters of an ECD represent the location, orientation, and strength of the current in the activated brain area. The ECDs were identified by searching for systematic local changes, persisting tens of milliseconds, in the measured magnetic field pattern. ECD model parameters were then determined at those time points at which the magnetic field pattern was clearly dipolar. The software identifies the sensor measuring the strongest signal at the channels covering the field pattern, and uses a location below this sensor as a seed point for the following ECD model parameter estimation. The parameter fit does not depend on the exact selection of the seed point in the local neighbourhood of the maximum signal. Only ECDs explaining more than 85% of the local field variance during each dipolar response peak were accepted in the multidipole model. Based on this criterion, 3–4 spatiotemporal components were selected into the individual subjects' models. The analysis was then extended to the entire time period, and all MEG channels were taken into account: The previously found ECDs were kept fixed in orientation and location while their strengths were allowed to change.

For optimizing the accuracy of the spatial fits, the orientation and location of the ECDs were estimated in each individual in the condition with the strongest signals in the time windows of the main experimental effects suggested by the sensor level data. However, the variability in the signal-to-noise ratios between conditions was very small, and, on the basis of visual inspection and on the calculated goodness-of-fit values obtained by comparing the original data and the data predicted by the fitted sources, the same sources explained well the responses in the other conditions.

Due to the variability of the response shape across individuals, the 250-ms response amplitudes were estimated as an average over a 50-ms (for ambiguous sounds) or 100-ms window (separately for congruent and incongruent conditions) around the individual response

peaks. For consistency, 100-ms response amplitudes were estimated from 50-ms time windows around the individual response peaks.

The ECD source waveforms (average strengths and peak latencies of the responses) were statistically tested using ANOVA and paired t tests (two-sided, Bonferroni corrected). Effect sizes μ were estimated as the difference between two condition means divided by a standard deviation of the data across both conditions.

Source analysis: Minimum Norm Estimates

In the auditory modality, ECD models have been shown to coincide well with distributed modelling approaches (Vartiainen et al., 2009). For verifying the spatial distribution of activity obtained with ECD modeling, the cortical generators were additionally visualized with a distributed source model, using MNE Suite software package (M. Hämäläinen, Martinos Center for Biomedical Imaging, Massachusetts General Hospital). MNE implements the L2 minimum norm estimate of the source distribution, which seeks for current distribution that explains the measurements and has the smallest L2-norm. MNE analysis results in distributed models of the cortical activation, but provides little information of the shape or extent of the activated area.

For MNE analysis, the cortical surface of each subject was reconstructed from the corresponding MR images with the Freesurfer software (Dale and Sereno, 1993; Fischl et al., 1999). Each hemisphere was covered with ~5000 potential source locations. Currents oriented normal to the cortical surface were favoured by weighting the transverse currents by a factor of 0.3 (Lin et al., 2006), and depth-weighting was used to reduce the bias towards superficial sources. Noise-normalized MNEs (dynamical Statistical Parametric Maps, dSPMs) were calculated over the whole cortical area to estimate the signal-to-noise ratios in each potential source location (Dale et al., 2000). Noise covariance matrix was estimated from the 200-ms prestimulus baseline periods in the raw data.

For group-level visualization, the MNEs of individual subjects were first normalized to the maximum value of that subject and subsequently morphed, with spatial smoothing, to one subject's brain. The statistical analysis of MNEs was performed, by means of paired two-sided t tests, on each subject's normalized values within a region of interest (ROI) centered around the Heschl's gyrus that contained both the MNE maxima and the ECD models of all subjects.

Results

Congruent vs. Incongruent Sounds: Sensor-level results

The initial sensor-level analysis revealed that all four stimuli within the stimulus blocks evoked strong responses bilaterally over the temporal areas, peaking at about 100 ms and at 250–700 ms after the onset of each sound. Figure 3 depicts the areal averages of the sensor-level signals (for the whole-head sensor-level data, see Figure 4). The 100-ms (N100m) re-

sponses were attenuated for the stimuli at positions 2nd–4th compared with the first stimulus, similarly in all conditions. An additional response at around 250 ms was observed in both incongruent conditions.

Congruent vs. Incongruent Sounds: Source-level results

Despite the careful acoustic matching of stimuli, the N100m responses to the first stimuli in a block were statistically significantly smaller for the cat than voice sounds in the left hemisphere (LH) as modelled by the ECDs (t test $p < 0.02$, effect size $\mu = 0.7$), whereas the N100m responses to other stimulus positions did not differ significantly between cat and voice stimuli in either hemisphere.

For the last stimulus, the incongruent sounds evoked prominent responses at ~250 ms after the stimulus onset in the right hemisphere (RH), without statistically significant differences between the cat and voice contexts (ECD analysis, Congruency x Category type interaction, $F1,7 = 0.64$; $p = 0.43$), suggesting that the effect was not specific to female voices nor cat vocalizations. Thus for the subsequent analysis of the congruent/incongruent sounds, the responses to cat and voice sounds were averaged together.

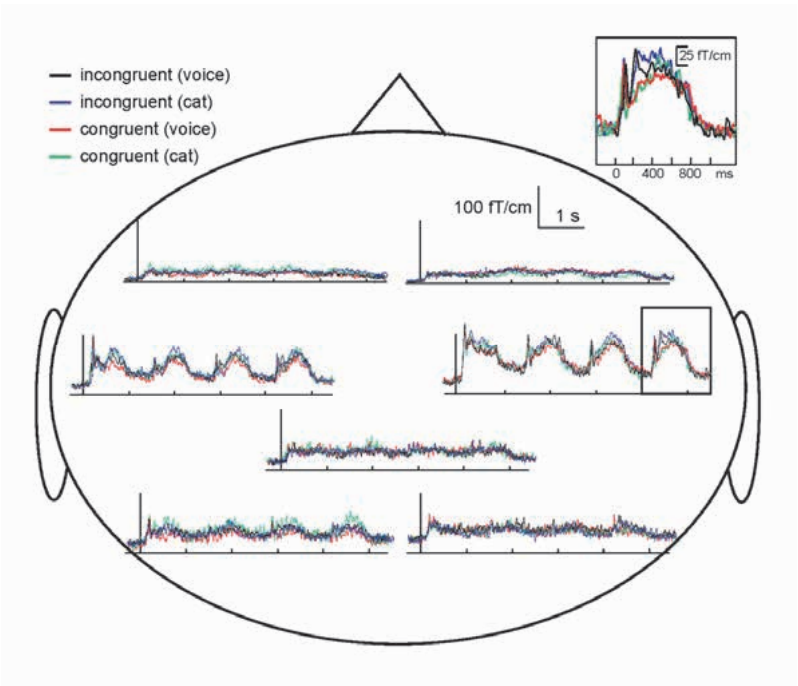

Figure 3. MEG signals. Areal responses over all subjects to congruent and incongruent trials. Incongruent (voice) refers to an experimental condition in which the three first sounds were voice sounds, and the last sound was a cat sound. The insert shows enlarged responses to the last sounds, recorded over the right temporal cortex.

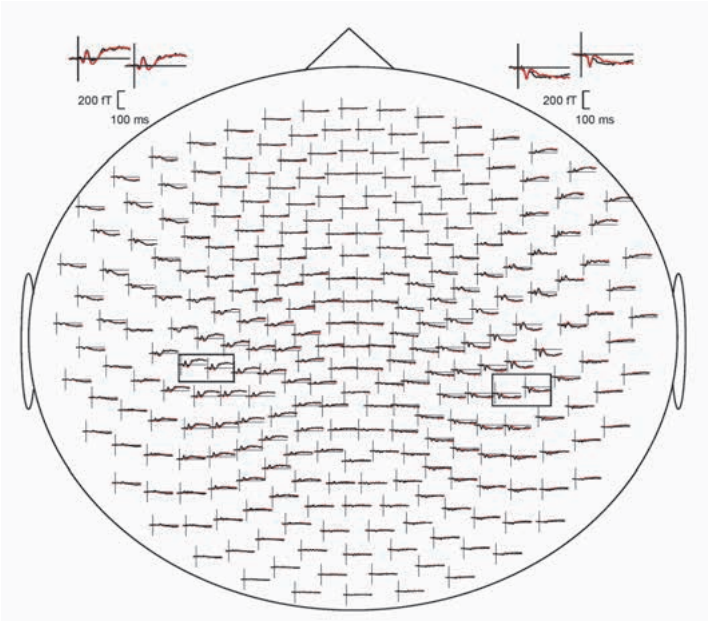

Figure 4. Responses at the 275 MEG channels averaged over all subjects for the incongruent (black) and congruent (red) conditions. The inserts depict the maximum channels over the left and right hemispheres.

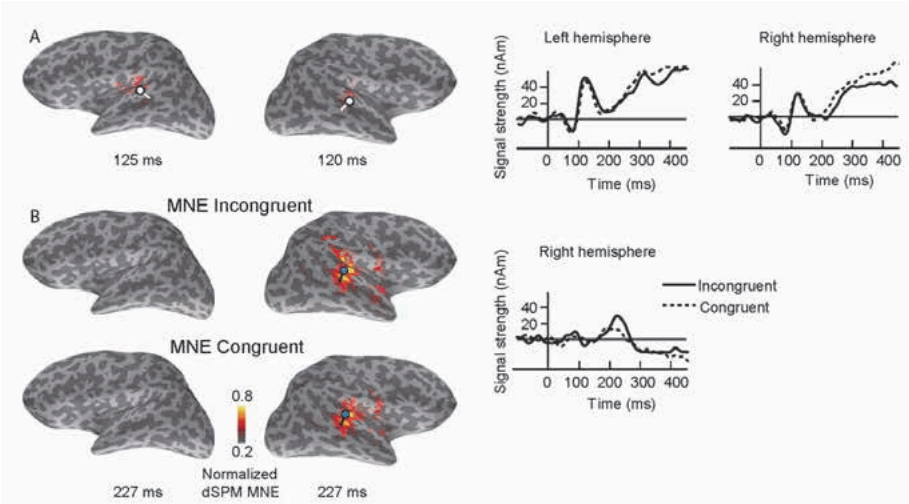

Figure 5. MEG source analysis in one subject. The locations (dots) and orientations (tails) of the ECDs used to model the N100m responses (A, white dots), and of the right-hemispheric 250-ms responses in incongruent and congruent conditions (B, blue dots) in one subject, superimposed on the subject's MNE dSPM distributions. The inserts (right) depict the corresponding ECD time courses in a time window of -100 ms to 450 ms with respect to the stimulus onset.

Figure 5 depicts the ECDs, the corresponding source waveforms, and the MNE dSPMs of one subject to the last sounds in the incongruent and congruent conditions, superimposed on her reconstructed cortical surface. In agreement with previous studies (for a review, see (Hari, 1990), the N100m responses were adequately explained by two ECDs, one in the left and one in the right supratemporal auditory cortex (indicated by white dipoles). The same sources explained also the sustained activity peaking > 300 ms. In the RH, another source with more supero-posterior location was needed to explain the responses around ~250 ms (indicated by a blue dipole). The ECD and MNE analyses suggested rather similar sequence of cortical activation: Both methods indicated right-hemispheric temporo-parietal activation ~230–250 ms that was stronger in the incongruent than congruent stimulus condition.

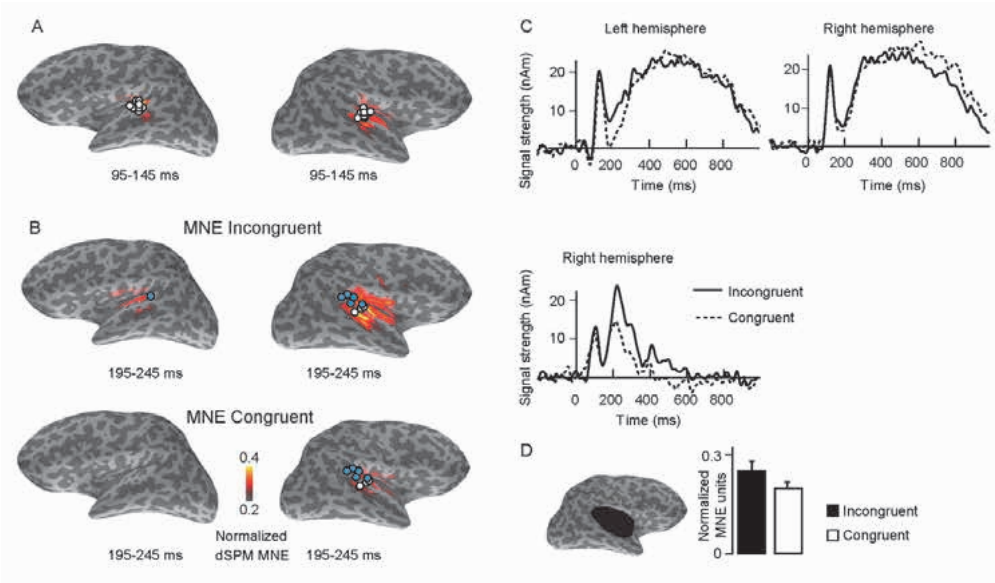

Figure 6. MEG group-level data. The locations of the ECDs used to model the N100m responses (white dots, A), and the 250-ms responses (blue dots) in congruent and incongruent conditions (B) in all subjects, morphed and superimposed on the average MNE dSPM distributions. Note that in three subjects, the same ECD was used to model both the N100m and the 250-ms response in the right hemisphere (white dots in B). C. ECD time courses from -200 ms to 1000 ms with respect to the stimulus onset. D MNE ROI analysis on the mean activation over the marked cortical area in the time window of 195–245 ms.

Figure 6 illustrates the ECDs, the corresponding source waveforms, and the MNE dSPMs over all subjects to the last sounds in the incongruent and congruent conditions, morphed and superimposed on one subject's reconstructed cortical surface.

The ECD models for the different subjects consisted typically of two ECDs in the RH, and one ECD in the LH (Fig. 6A). In three subjects, ECDs explaining the field patterns around 100 ms and 250 ms in the RH were located close to each other and had very similar orientations, and to prevent interactions between these ECDs, the same ECD was used to model both responses. In one subject, a 4th ECD was needed in the LH to explain the magnetic field variations at ~250 ms (Fig. 6B). While the N100m responses were consistently located in the vicinity of planum temporale in both hemispheres in all subjects, the location of the 250-ms responses showed more interindividual variability.

The N100m responses peaked in the LH at 108 ± 8 ms and at 113 ± 7 ms (mean ± SEM), respectively, in the incongruent and congruent conditions, and in the RH at 113 ± 5 ms in both conditions, without significant differences in the ECD peak latencies or mean response amplitudes between conditions. At the LH, the responses at ~200 ms explained by the same ECDs tended to be stronger for incongruent than congruent sounds, but this difference did not reach statistical significance (estimated individually from a 50-ms time window around the maximum difference between conditions, t test $p = 0.15$).

The RH 250-ms responses peaked at 230 ± 10 ms in the incongruent condition, and at 231 ± 12 ms in the congruent condition. The responses were statistically significantly stronger for the incongruent than congruent sounds as modelled by the ECDs (estimated from a 100-ms time window around the individual peak responses, t test $p < 0.01$, effect size $\mu = 0.9$; For individual source waveforms, see Figure 7). ROI analysis of the maximum MNE maps over the right temporo-parietal region gave consistent results (average over the time window of 195–245 ms in the incongruent vs. congruent conditions, t test $p < 0.03$, effect size $\mu = 0.7$; Fig. 5D).

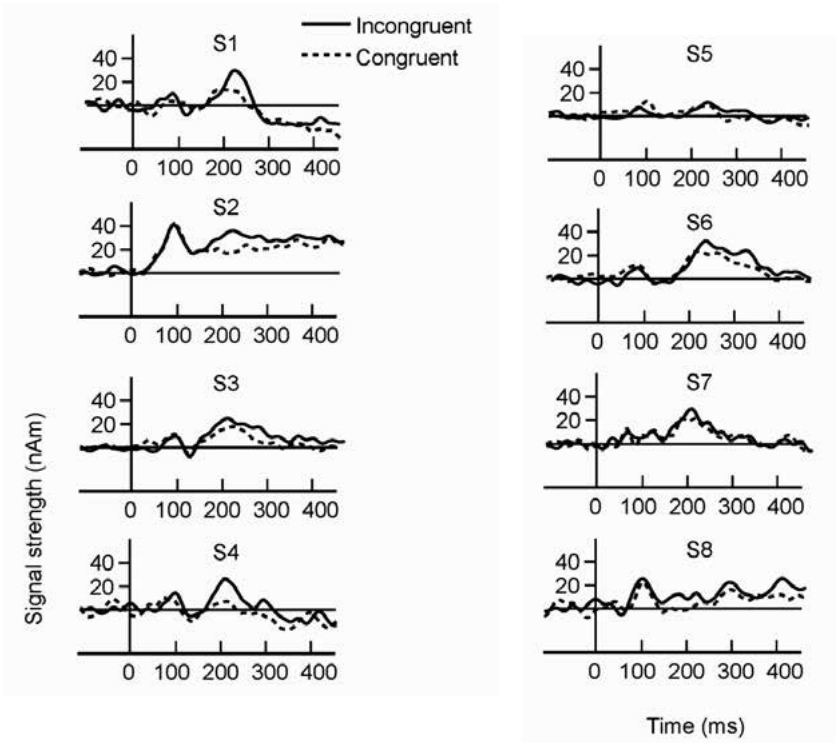

Figure 7. The individual source waveforms of ECDs used to model the 250-ms responses in the right hemisphere (subjects S1-S8).

"Ambiguous" Sounds

For testing the categorical adaptation effects in two different contexts ("voice" and "cat"), we used acoustically identical target sounds that were derived from the voice sounds (see Methods). Whereas the N100m responses to these ambiguous sounds presented after cat and voice stimuli did not differ from each other, the right-hemispheric responses peaking at 265 ± 28 ms were statistically significantly stronger to the target sounds presented after the voice than cat stimuli as modelled by the ECDs (estimated from a 50-ms time window around the individual peak responses, t test $p < 0.02$, effect size $\mu = 1.1$; see Fig. 8).

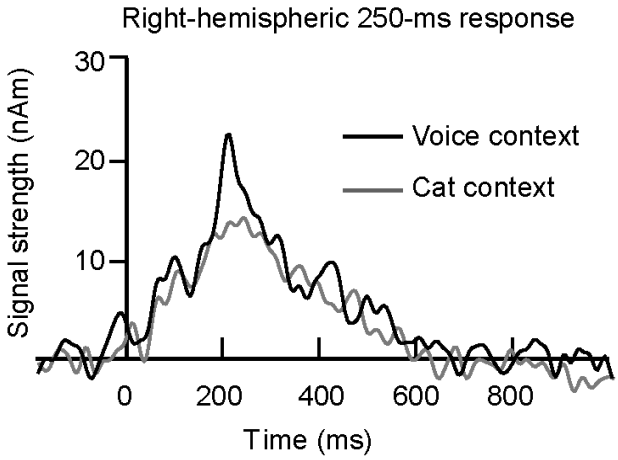

Figure 8. The mean time courses of the right-hemispheric 250-ms responses for ambiguous sounds in cat and voice contexts, averaged over all subjects.

Discussion

In the present study, we investigated the temporal processing of auditory categories by utilizing carefully-matched human and cat vocalizations. In particular, we used a paradigm that enabled us to compare responses to physically identical stimuli presented in different categorical contexts. Our results demonstrate that, when the low-level auditory stimulus differences are minimized, responses specifically at the right temporoparietal cortex react vigorously to auditory categorical violation regardless of the stimulus category at ~200–250 ms after the stimulus onset.

Although our experimental stimuli were matched for several temporospectral acoustic characteristics, for the easily-recognizable stimuli, the overall harmonic structures still differed enough to provide cues needed for successful online categorization of the sounds. The conspicuous auditory N100m responses can be evoked by any sound onset or change in the auditory environment, but they also indicate stimulus-specific neural activity (Hari, 1990). Indeed, in the left hemisphere the first N100m responses for a stimulus block were stronger for voice than cat sounds, probably reflecting the remaining acoustic differences between the sounds. This effect may be partly explained by the female voice stimuli containing more energy at the frequency level of 1000–1500 Hz than the cat vocalizations, although effect of stimulus bandwidth on cortical responses has been shown to be highly stimulus specific (Seither-Preisler et al., 2003; Shahin et al., 2005; Soeta et al., 2005). Thus, the use of a paradigm that allowed us to present the exact same stimuli in different

categorical contexts can be considered crucial for the interpretation of the results. The dif-differences between the congruent and incongruent sounds at ~200–250 ms after sound onset in the right hemisphere, present regardless of the sound category, suggest that at this time window, auditory processing has proceeded to a stage at which categorical templates have been established. Previously, right-lateralized auditory cortical fMRI activation in response to species-specific vocalizations has been reported in humans and monkeys, mainly in the STG/STS region (Belin et al., 2002; Belin and Zatorre, 2003; Formisano et al., 2008; Petkov et al., 2008), and right-hemispheric STG/STS has recently been related to speaker-related changes in pitch that are needed for recognizing speech among changing speakers (von Kriegstein et al., 2010). Several earlier neuroimaging studies have pointed to functional asymmetries in the auditory areas, with the left and right auditory cortices being predominantly sensitive to temporal and spectral changes, respectively (e.g., (Obleser et al., 2008; Zatorre and Belin, 2001). Our MEG results for categorizing vocalizations—for which rapid analysis of spectral information is crucial—are in agreement with these results, and further suggest the observed activity to support categorical processing at ~200–250 ms after sound onset.

Recently, human vocalizations were demonstrated to evoke stronger responses than animal vocalizations at 169–219 ms after sound onset within the anterior right STG and STS, without topographical differences between stimuli (De Lucia et al., 2010). The current data suggest that, after rather strict stimulus control for both acoustical features and attentional demands, auditory MEG responses to human voices and cat sounds did not statistically significantly differ from each other at around 200–250 ms. Rather, our results suggest the right posterior temporoparietal cortex to be especially activated in response to auditory categorical violation, regardless of the actual auditory stimulus. In 5 out of 8 subjects, the source for this response was separable from the source of the N100m response that has repeatedly been localized to the posterior part of the planum temporale (Hari, 1990). However, taken the relatively large interindividual variability in the 250-ms response source locations, they are likely to reflect anatomically more widespread synchronous activity, possibly including also the planum temporale that has earlier been suggested to be engaged in segregating and matching spectrotemporal patterns crucial for auditory object recognition (Griffiths and Warren, 2002). Combining electrophysiological measures e.g., with functional MRI could in the future provide more detailed spatial information on these responses.

The 250-ms responses in the present study had a fairly similar polarity to the N100m responses, and their cortical sources were located at the near vicinity of those of N100m with right-hemisphere dominance. These sources are unlikely to reflect the well-established, broad positive component at ~300 ms (P300) evoked by infrequent task-relevant stimuli in EEG recordings, likely reflecting widespread activity with bilateral sources at occipito-temporal, centro-temporal, parietal and precuneal areas (Anurova et al., 2005). Rather, our 250-ms responses seem to overlap temporally and spatially with activity that has been

observed, although bilaterally, in earlier auditory MEG studies on processing syllables, spo-spoken words, and environmental sounds (Bonte et al., 2006; Renvall et al., 2012; Uusvuori et al., 2008). These responses do not seem to react to, e.g., phonetic or semantic task manipulations (Bonte et al., 2006; Uusvuori et al., 2008). Future studies are needed to explore whether these responses are related e.g., to accessing templates for different auditory categories regardless of stimulus type, possibly with different hemispheric emphasis for speech-like sounds.

The careful stimulus control can also be considered the main limitation of our present study: The stimuli were simple and they were constructed as continua from two exemplars. Even though their variability was increased by filtering and transposing them to different pitches, their ecological validity remains limited, compared with e.g., spoken words or environmental sounds. In future studies, the representation of auditory categories should be addressed also using more realistic auditory scenes, for example by modifying stimulus recognizability with varying level of superimposed noise (Renvall et al., 2012) and using a wider range of stimulus categories.

Although at the behavioral level the categorical context did not statistically significantly affect the categorization of ambiguous sound stimuli, the cortical responses to these sounds differed greatly depending whether they were presented after cat or voice sounds. Specifically, the right-hemispheric 250-ms responses were statistically significantly greater to sounds presented in the voice than cat context although the ambiguous sounds were acoustically closer to the voice stimuli. This finding could suggest that human voices as potentially more meaningful stimuli for the listener generated a stronger contextual effect, and thus resulted in a greater categorical mismatch for sounds that could not be un-ambiguously attributed to one of the two categories. This suggests a more established status for processing of human voices in the human auditory cortex than e.g. animal vocalizations (Fecteau et al., 2004). However, further studies are evidently needed for establishing the complex interactions between context and target sounds. Specifically if the target sounds such as the ambiguous sounds here do not belong to any natural category, different cortical mechanisms may also apply.

In conclusion, our present results suggest that, after careful matching of acoustic stimulus features and behavioral demands, auditory categories for vocalizations are accessed by ~250 ms, preferably in the right posterotemporal cortex. This activity may reflect the detailed spectral analysis needed in the auditory categorical distinction of vocalizations.

Acknowledgements

We thank Niclas Kilian-Hütten and Jasper van den Bosch for help with the behavioral measurements, Mia Illman for the surface reconstructions, and Jan Kujala, Miiamaaria Kujala, Lauri Parkkonen and Tiina Parviainen for comments on the manuscript. This work was

supported by the Academy of Finland (National Centers of Excellence Programme 2006-2011, and grant numbers #213828 and 127401 to HR), Netherlands Organisation for Scientific Research, Helsingin Sanomat Centennial Foundation, Emil Aaltonen Foundation and The Ella and Georg Ehrnrooth Foundation.

References

Alain, C., Arnott, S., Hevenor, S., Graham, S., and Grady, C. (2001). "What" and "where" in the human auditory system. Proc Natl Acad Sci USA *98*, 12301−12306.

Altmann, C., Nakata, H., Noguchi, Y., Inui, K., Hoshiyama, M., Kaneoke, Y., and Kakigi, R. (2008). Temporal dynamics of adaptation to natural sounds in the human auditory cortex. Cereb Cortex *18*, 1350−1360.

Anurova, I., Artchakov, D., Korvenoja, A., Ilmoniemi, R. J., Aronen, H. J., and Carlson, S. (2005). Cortical generators of slow evoked responses elicited by spatial and nonspatial auditory working memory tasks. Clin Neurophysiol *116*, 1644−1654.

Belin, P., Fecteau, S., and Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. Trends Cogn Sci *8*, 129−135.

Belin, P., Zatorre, R., and Ahad, P. (2002). Human temporal-lobe response to vocal sounds. Cognitive Brain Research *13*, 17–26.

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

Belin, P., and Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport *14*, 2105– 2109.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonic-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences *17*, 97-110.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International *5*, 341−345.

Bonte, M., Parviainen, T., Hytönen, K., and Salmelin, R. (2006). Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. Cereb Cortex *16*, 115–123.

Charest, I., Pernet, C., Rousselet, G., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J., and P., B. (2009). Electrophysiological evidence for an early processing of human voices. BMC Neurosci *10*, 127.

Dale, A., and Sereno, I. (1993). Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: A linear approach. J Cogn Neurosci *5*, 162-176.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., and Halgren, E. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. Neuron *26*, 55-67.

De Lucia, M., Clarke, S., and Murray, M. (2010). A temporal hierarchy for conspecific vocalization discrimination in humans. J Neurosci *30*, 11210-11221.

Fecteau, S., Armony, J., Joanette, Y., and Belin, P. (2004). Is voice processing species-specific in human auditory cortex? An fMRI study Neuroimage *23*, 840-848.

Fischl, B., Sereno, M., Tootell, R., and Dale, A. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum Brain Mapp *8*, 272-284.

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? Brainbased decoding of human voice and speech. Science *322*, 970-973.

Friston, K. (2005). A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci *360*, 815-836.

Garrido, M. I., Friston, K. J., Kiebel, S. J., Stephan, K. E., Baldeweg, T., and Kilner, J. M. (2008). The functional anatomy of the MMN: a DCM study of the roving paradigm. Neuroimage *42*, 936-944.

Griffiths, T., and Warren, J. (2002). The planum temporale as a computational hub. Trends in Neurosciences *25*, 348−353.

Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. (1993). Magnetoencephalography – theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev Mod Phys *65*, 413–497.

Hansen, P. C., Kringelbach, M. L., and Salmelin, R., eds. (2010). MEG - An introduction to methods (New York, Oxford UP).

Hari, R. (1990). The neuromagnetic method in the study of the human auditory cortex. In Auditory Evoked Magnetic Fields and Electric Potentials, F. Grandori, M. Hoke, and R. G.L., eds. (Basel, Karger), pp. 222–282.

Jaaskelainen, I. P., Ahveninen, J., Bonmassar, G., Dale, A. M., Ilmoniemi, R. J., Levanen, S., Lin, F. H., May, P., Melcher, J., Stufflebeam, S*., et al.* (2004). Human posterior auditory cortex gates novel sounds to consciousness. Proc Natl Acad Sci U S A *101*, 6809-6814.

Kaas, J., and Hackett, T. (1999). 'What' and 'where' processing in auditory cortex. Nature Neuroscience *2*, 1045−1047.

Lewis, J. W., Brefczynski, J. A., Phinney, R. E., Janik, J. J., and DeYoe, E. A. (2005). Distinct cortical pathways for processing tool versus animal sounds. J Neurosci *25*, 5148–5158.

Lin, F., Witzel, T., Ahlfors, S., Stufflebeam, S., Belliveau, J., and Hämäläinen, M. (2006). Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. Neuroimage *31*, 160-171.

Medendorp, W., Kramer, G., Jensen, O., Oostenveld, R., Schoffelen, J., and Fries, P. (2007). Oscillatory activity in human parietal and occipital cortex shows hemispheric lateralization and memory effects in a delayed double-step saccade task. Cereb Cortex *17*, 2364-2374.

Murray, M., Camen, C., Gonzalez Andino, S., Bovet, P., and Clarke, S. (2006). Rapid brain discrimination of sounds of objects. J Neurosci *26*, 1293−1302.

Näätänen, R. (1992). Attention and brain function (Hillsdale, NJ, Lawrence Erlbaum Associates).

Obleser, J., Eisner, F., and Kotz, S. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. J Neurosci *28*, 8116–8124.

Petkov, C., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., and Logothetis, N. (2008). A voice region in the monkey brain. Nature Neuroscience *11*, 367-374.

Rauschecker, J., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc Natl Acad Sci USA *97*, 118001−111806.

Renvall, H., Formisano, E., Parviainen, T., Bonte, M., Vihla, M., and Salmelin, R. (2012). Parametric merging of MEG and fMRI reveals spatiotemporal differences in cortical processing of spoken words and environmental sounds in background noise. Cereb Cortex *22*, 132-143.

Romanski, L., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P., and Rauschecker, J. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nature Neuroscience *2*, 1131−1136.

Salmelin, R., Hari, R., Lounasmaa, O. V., and Sams, M. (1994). Dynamics of brain activation during picture naming. Nature *368*, 463−465.

Seither-Preisler, A., Krumbholz, K., and Lutkenhoner, B. (2003). Sensitivity of the neuromagnetic N100m deflection to spectral bandwidth: a function of the auditory periphery? Audiol Neurootol *8*, 322-337.

Shahin, A., Roberts, L. E., Pantev, C., Trainor, L. J., and Ross, B. (2005). Modulation of P2 auditory-evoked responses by the spectral complexity of musical sounds. Neuroreport *16*, 1781-1785.

Soeta, Y., Nakagawa, S., and Tonoike, M. (2005). Auditory evoked magnetic fields in relation to bandwidth variations of bandpass noise. Hear Res *202*, 47-54.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., and Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. Curr Biol *19*, 498−502.

Uusvuori, J., Parviainen, T., Inkinen, M., and Salmelin, R. (2008). Spatiotemporal interaction between sound form and meaning during spoken word perception. Cereb Cortex *18*, 456–466.

Vartiainen, J., Parviainen, T., and Salmelin, R. (2009). Spatiotemporal convergence of semantic processing in reading and speech perception. J Neurosci *29*, 9271-9280.

von Kriegstein, K., Smith, D., Patterson, R., Kiebel, S., and Griffiths, T. (2010). How the human brain recognizes speech in the context of

changing speakers. Journal of Neuroscience *30*, 629−638.

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., and Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. Proc Natl Acad Sci U S A *108*, 20754-20759.

Warren, J., and Griffiths, T. (2003). Distinct mechanisms for processing spatial sequences and pitch sequences in the human auditory brain. J Neurosci *23*, 5799−5804.

Warren, J., Scott, S., Price, C., and Griffiths, T. (2006). Human brain mechanisms for the early analysis of voices. Neuroimage *31*, 1389−1397.

Zatorre, R. J., and Belin, P. (2001). Spectral and temporal processing in human auditory cortex. Cereb Cortex *11*, 946–953.

CHAPTER 4

Cortical processing of spatial cues in natural auditory scenes

Corresponding publication: Staeren, N., Renvall, H., Schreiner, C., Walter, A., Goebel, R., Formisano, E. (in preparation). Cortical processing of spatial cues in natural auditory scenes.

Summary

The segregation of an auditory object from a sound mixture (or auditory scene) requires the interplay between bottom-up processing of the acoustic scene elements and top-down processes of attentive selection and binding. Spatial hearing contributes to this analysis by providing cues on location and motion of the sound sources. This study investigates the cortical processing of spatial cues during listening of natural auditory scenes. Using the technique of binaural recording and in-ear microphones, we recorded realistic auditory scenes containing two concurrent sound sources, a voice centrally located in front of the listener (foreground), and an environmental sound located at different locations at the background. During fMRI measurements subjects were instructed to attend one of the sound sources ("*Voice*" vs "*Environment*"), under two distinct playback conditions: 1) Stereo playback which preserves the spatial acoustic information of the original recordings ("*Spatial*") or 2) Mono playback, which removes spatial information ("*Non*-spatial"). Our analyses show that processing of the spatial cues - independently of the attention condition - corresponded with significantly increased brain activation at the bilateral posterior superior temporal areas. These regions are known for processing spatial and motion information ("where" stream). However, we also observed significant activation differences in the *Spatial* vs *Non-spatial* comparison that depended on the attention target. When listeners attended to environmental background sounds, we found significant differences in left planum temporale and left inferior frontal gyrus. Conversely, when listeners attended to vocal sounds, we found significant activation differences in bilateral clusters of middle superior temporal gyrus and sulcus, which overlap with voice sensitive regions. These attention-dependent effects suggest that – in order to segregate an auditory source from a sound mixture - spatial cues are integrated with other relevant spectral and temporal cues in cortical locations specifically involved in the recognition of sounds.

Introduction

A natural environment rarely contains one sound. Overlapping voices, mechanical background noise, a phone ringing; in most cases the acoustic signal at our ears comprises sounds from several sources. Automatic and effortless for most of us, segregation of the sources from a complex sound mixture (or auditory scene), is a formidable example of the computational capabilities of our auditory system. Processing of a scene into perceptual auditory objects is determined by the interplay between bottom-up processing of the spectral and temporal relations of the acoustic scene elements and top-down processes of attentive selection and enhancement of the relevant sounds (Bregman, 1990).

Spatial hearing also contributes to the processing of auditory scenes by providing information on location and motion of the sound sources. As there is no explicit representation of auditory space on the receptor surface, the auditory system derives the information on the location and motion of the sources from various acoustic cues. Locations in the vertical plane and in the front-back direction are resolved from the direction-dependent modifications of spectral profile generated by the outer ear and the head (spectral cues). Horizontal localization of the sound sources relies on timing and level differences at the two ears (interaural timing [ITD] and level [ILD] difference). Perception of sound motion relies on the analysis of dynamic changes of these cues.

The neural analysis of spatial acoustic cues starts in the brain stem at level of the superior olivary complex (SOC). At the level of the inferior colliculus (IC), all the individual spatial acoustic cues have been processed and filtered (Groh et al., 2003). These separate cues are then integrated in the next synaptic levels of the thalamo-cortical system. In the cortex, the location of a sound source is represented by populations of distributed and broadly tuned neurons (Recanzone et al., 2000). When comparing the spatial selectivity of neurons in different fields of the auditory cortex in the macaque, a sharper spatial tuning is found in caudal fields (CM (Tian et al., 2001), (Recanzone et al., 2000) and CL (Recanzone et al., 2000), (Miller and Recanzone, 2009)) compared to A1 or to antero-lateral fields. Furthermore, deactivation of the posterior auditory field, in the cat, causes behavioral dysfunctions in sound localization (Lomber and Malhotra, 2008). These results provide strong support to the proposal of a dorsal ('where') stream of auditory areas specialized for the processing of spatial information. Anatomical studies indicate that extensive connections exist between these caudal auditory fields and spatial domains of the prefrontal cortex (Romanski et al., 1999).

Results from neuroimaging studies in humans are generally supportive of this hypothesis. Several studies investigating the cortical basis of sound localization (Alain et al., 2001; Altmann et al., 2007; Barrett and Hall, 2006; Warren and Griffiths, 2003) and motion (Baumgart et al., 1999; Hart et al., 2004; Krumbholz et al., 2005a; Krumbholz et al., 2005b; Pavani et al., 2002; Warren et al., 2002; Warren et al., 2005) have reported a selective activation of posterior temporal regions (planum temporale, [PT] and posterior superior

temporal gyrus [pSTG]), and of regions at the temporal – parietal boundaries (Lewis et al., 2000). Activation of these regions appears to be prominent when subjects are actively involved in a task of sound localization (Zatorre et al., 2002) and in the presence of sound motion (Getzmann and Lewald, 2010; Warren et al., 2002).

However, this interpretation of posterior temporal activation in terms of functional specialization for spatial audition is not univocal. For example, it has been suggested that the activation of the PT does not reflect spatial processing per se but rather the integration of spatial information with auditory object information (Zatorre et al., 2002). This alternative interpretation is supported by the findings that manipulation of the number of auditory objects in a scene produces effects in PT similar to spatial manipulations (Smith et al., 2010).

In the present study we examined the cortical processing of spatial cues embedded in realistic auditory scenes. Using ear-insert microphones, we recorded a set of naturalistic scenes that contained a vocal sound centrally located in front of the listener and an environmental sound located at the peripheral background (e.g. a voice with a car passing). During functional MRI (fMRI) measurements, subjects attentively listened to the auditory scenes, under two distinct playback conditions: 1) Stereo playback ("*Spatial*") which preserves the spatial acoustic information of the original recordings (e.g. motion of the sound on the background) or 2) Mono playback, which removes spatial information ("*Non-spatial*"). Furthermore, we manipulated the top-down context for processing the auditory scenes by directing the subjects' attention either to the voice in the foreground or to the background sounds ("*Voice*" vs "*Environment*"). This design allowed us to examine the relation between mechanisms for the analysis of spatial cues and attention mechanisms responsible for selecting and segregating sound objects from a scene. In particular, we aimed at distinguishing cortical regions involved in the automatic (i.e. attention-independent) analysis of spatial cues from regions involved in integrating spatial and sound object information during auditory scene analysis.

Experimental Procedures

Subjects

We studied, with informed consent, 10 adults (mean age ± SD: 28 ± 4 yrs; 4 females, 6 males, one left-handed). All subjects were graduate university students and were paid for their participation. Subjects had no history of hearing or neurological impairments, and were naive to the experimental setup. The study received a prior approval by the Ethical Committee of the Faculty of Psychology, Maastricht University.

Stimuli

Eighty auditory scenes were created by using excerpts from audio recordings from 12 vocal actors and 30 environments. Sounds were recorded binaurally using two in-ear microphones (FG-23652-P16, Knowles Electronics, Itasca, Illinois, U.S.A.) and a portable digital recorder

(96 Khz, 24bit, M-Audio MicroTrack 24/96 Pocket Digital Recorder). After recording, sounds were down-sampled to 44.1 KHz/16 bit using Adobe Audition (Adobe Systems, Inc., CA, USA). The duration of the sounds was between 450 and 2635 ms (mean length ± SD: 1306 ± 565 ms); amplitude envelopes and average root-mean-square levels of the sounds were matched using MATLAB 7.0.1 (The MathWorks, Inc., Natick, MA, USA).

Auditory scenes for the "Spatial" condition were created by mixing separately the two audio channels; a monaural version of the same scenes was created by merging the two audio channels. All stimuli in this study were recorded inserting the microphones in the ear canal of two listeners that did not take part in the fMRI measurements, and were played to the subjects via the MR-compatible headphones (see below). It is known that - because of inter-individual differences in head and external ear shape – non-individualized recordings as used in this study do not produce the same perceptual quality as individualized recordings. However, we choose not to record the stimuli individually because of the difficulty of recreating natural complex scenes for each subject. We assessed the quality of spatial perception in behavioral pre-tests. All listeners that participated in the fMRI measurements reported a clear spatial perception of our auditory stimuli outside and inside the MR scanner.

fMRI experimental design

A 2 x 2 block design with space ("Spatial" vs "Non-Spatial") and attention ("Voice" vs "Environment") as factors was used. The experiment consisted of 2 functional runs during which auditory scenes in the four different conditions were presented according to a block design. Each of the two runs (22 min/run) included 9 blocks per condition and four target blocks (see below); the sequence of conditions was randomized and blocks were separated by a fixation period of three TRs. Each block consisted of four TRs (TR= 4640 ms, total = 18.5 s) and an auditory scene was presented for each trial. Every block was preceded by a cue presented at the fixation point indicating the attention condition ("E" or "V"). Subjects were instructed to respond with a button press in case the attended sound in two consecutive auditory scenes was the same. This occurred in 10% of the cases ("target blocks"); there were 2 target blocks per conditions (4 blocks/run). The response hand was alternated across subjects.

fMRI measurements

Brain imaging was performed with a 3 Tesla Siemens Allegra (head setup) at the Maastricht Brain Imaging Center. In each subject, two runs of 282 volumes were acquired with a T2*-weighted gradient-echo planar imaging (EPI) sequence (TR = 4640 ms, voxel size = 2,5 $\times$ 2,5 $\times$ 2,5 mm^3, TE = 30 ms, FOV 256 $\times$ 256; matrix size 96 $\times$ 96, 32 slices covering the cortex). Anatomical images (1 $\times$ 1 $\times$ 1 mm^3) were collected between the two functional runs using a 3D-MPRAGE T1-weighted sequence. During the measurements, the stimuli were delivered binaurally via MR compatible headphones (Commander XG, Resonance Technology,

Northridge, CA) at a comfortable listening level. To minimize the effect of scanner noise, the sounds were presented during silent periods using a clustered volume EPI technique that allowed for presentation of auditory stimuli in silence between subsequent volume acquisitions (Riecke et al., 2007; van Atteveldt et al., 2004).

fMRI Data Analysis: pre-processing and univariate statistics

Functional and anatomical images were analyzed with BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). Preprocessing consisted of slice scan-time correction (using sinc interpolation), linear trend removal, temporal high-pass filtering to remove nonlinear drifts of seven or less cycles per time course, and 3-dimensional motion correction. Temporal low pass filtering was performed using a Gaussian kernel with FWHM of two data points. Functional slices were co-registered to the anatomical data, and both data were normalized to Talairach space (Talairach and Tournoux, 1988).

Statistical analysis of the fMRI data was based on voxel-by-voxel general linear modeling (GLM) of the time series. For each subject, a design matrix was formed using a predictor for each experimental condition ("Spatial-Voice", "Spatial-Environment", "Non Spatial-Voice", "Non Spatial-Environment") and for the target blocks. The predicted time courses were adjusted for the hemodynamic response delay by convolution with a canonical hemodynamic response function (sum of two gamma functions).

Cortex-based realignment was performed for aligning the functional time series of individual subjects and to perform random effect group-based statistics (Goebel et al., 2006). Statistical maps were thresholded and corrected for multiple comparisons (alpha = 0.05) on the basis of cluster-level statistical threshold estimation performed on the cortical surface data (Forman et al., 1995; Goebel et al., 2006).

Results

Listening to auditory scenes induced extensive activations of the superior temporal cortex bilaterally, including the Heschl's gyrus and surrounding regions on the superior temporal gyrus and sulcus (see Figure 1a). Additional activation was found in the left middle temporal gyrus (MTG), left inferior frontal gyrus (IFG) and bilateral inferior parietal lobule (IPL). This overall activation pattern was largely common to both the "Spatial" and "Non spatial" conditions.

"Spatial" vs "Non Spatial" scenes

To examine the brain regions involved in the processing of spatial cues we first compared the activation to "Spatial" vs "Non Spatial" scenes grouped across attention conditions. We observed significantly higher BOLD responses for the "Spatial" condition (see Figure 1b) bilaterally in the posterior STG regions. In the left hemisphere, this region was located at the

adjacency with the temporal-parietal border. In the right hemisphere, an additional cluster was present along the STS.

We further dissected the "Spatial" vs "Non Spatial" contrast by analyzing the two attention conditions separately, i.e. we performed the two orthogonal contrasts "Spatial-Environment" vs "NonSpatial-Environment" (see blue map in Figure 1c) and "Spatial-Voices" vs "NonSpatial-Voices" (see red map in Figure 1c).

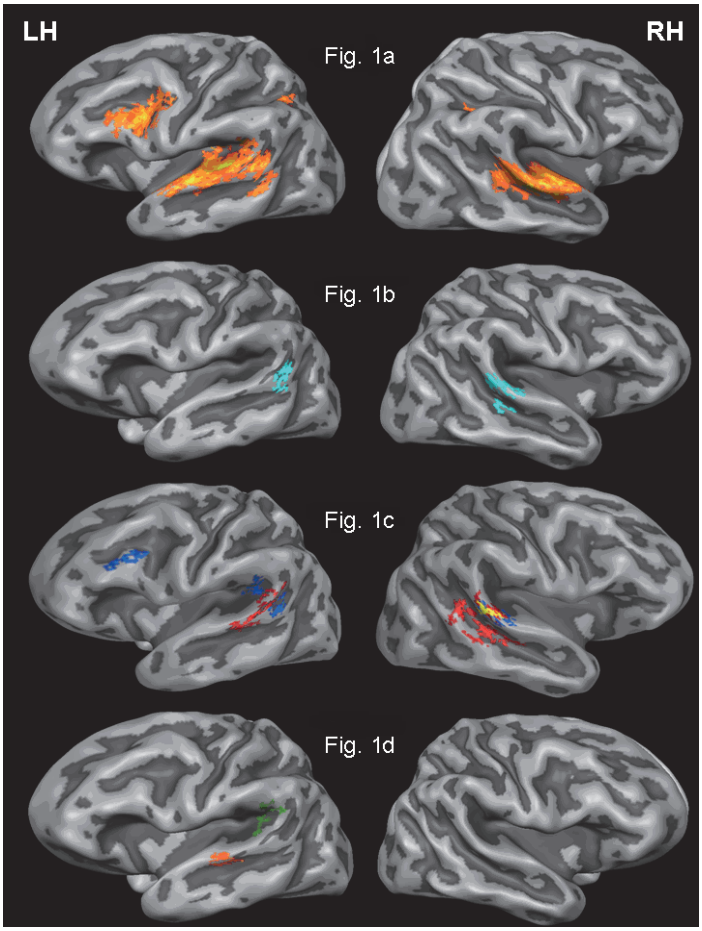

Figure 1: Results from Cortex-based Aligned Random Effect analysis using the General linear model. a) Overall activation map (SpVo + SpEn + NsVo + NsEn > baseline, F-map), b) Spatial versus Non-Spatial stimuli (SpVo + SpEn > NsVo + NsEn, c) Spatial versus Non-Spatial stimuli for either the Voice condition (Red, SpVo > NsVo) or the Environment condition (Blue, SpEn > NsEn). Voxels where both contrasts are significant (conjunction) are highlighted in yellow, d) Interaction maps: Green: SpEn – MoEn > SpVo – MoVo; Orange: SpVo – MoVo > SpEn - MoEn.

In the right posterior STG, we found a cluster of activation where these two contrasts were independently significant (see yellow map in Figure 1c and averaged time courses in Figure 2a). A similar cluster was also present in the left hemisphere, although it did not survive the corrected threshold (see also time courses in Figure 2b).

Besides these common clusters, activation clusters were detected specific to the different attention targets. When listeners attended to environmental background sounds, significant activation differences were found in the left planum temporale and in left inferior frontal gyrus (see blue color in Figure 1c). In these regions there was no activation difference for the orthogonal contrast ("Spatial-Voice" vs "NonSpatial-Voice"; see time courses in Figure 2c and 2d).

Conversely, when listeners attended to vocal sounds, we found significant activation differences in bilateral clusters of middle STG (left hemisphere) and STS (posterior and middle, right hemisphere). In these clusters - that resembled regions reported to be selectively activated for voices in previous studies (Belin et al., 2000) – there was no activation difference for the orthogonal contrast ("Spatial-Environment" vs "NonSpatial-Environment"; see time course in Figure 2e and 2f).

To test these observations statistically we calculated interaction maps, which are shown in Figure 1d. In these maps, of all the regions for which an individual contrast was significant (e.g. blue or red regions in Figure 1b) only the regions in the left PT and in the left middle STG survived a rigorous threshold ($p < 0.05$, corrected).

Attention to "Environment" vs attention to "Voice"

To examine the brain regions affected by the attention manipulation we compared the activation to the scenes grouped across spatial conditions (i.e."Environment" vs "Voice"). We observed significantly higher BOLD responses for the "Environment" condition (see Figure 3a) in a largely left-lateralized network of regions including posterior STG, posterior STS/MTG and, in the frontal lobe, and the dorsolateral prefrontal cortex (DLPFC). Bilateral activation of the posterior parietal cortex (PPC) and the left precentral gyrus (PrG) were also observed. No region showed increased activation for "Voice" compared to "Environment". When analyzing the two spatial conditions separately, ("Spatial-Environment" vs "Spatial-Voice" (see purple map in Figure 3b) and "Non Spatial-Environment" vs "Non Spatial-Voices" (see green map in Figure 3b), we observed a similar pattern of overall activation. Interestingly, however, there was little overlap between the two maps in frontal and parietal regions and only a common cluster of activation in left STS/MTG (see yellow map in Figure 3b and time course in Figure 3c).

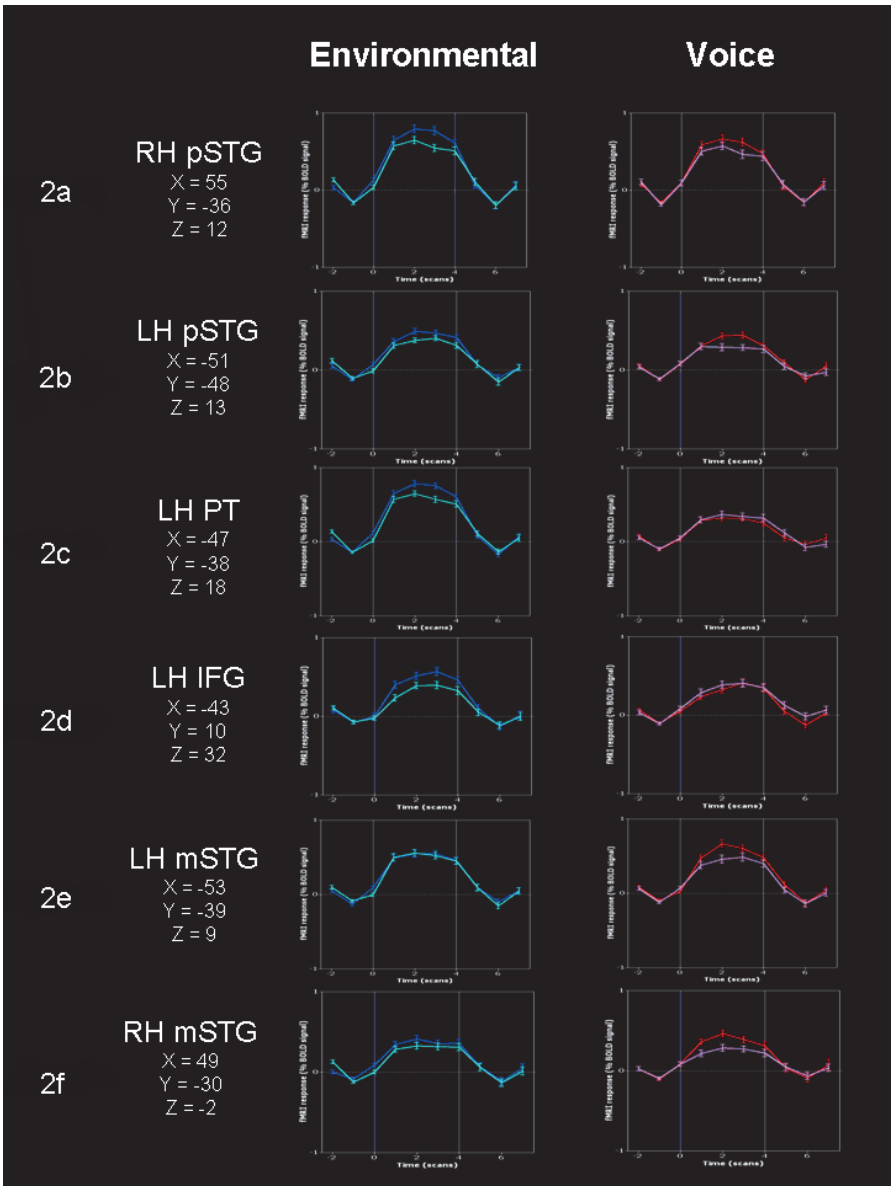

Figure 2: Average time courses for relevant clusters of the maps in Figure 1b-1d. The averaged brain activation during the "Spatial" conditions (dark blue and red lines) are compared to the brain activation in the "Non-Spatial" conditions (lighter blue and pink lines) separately for the attention to "Environment" (left column, blue) or to "Voice" (right column, red) condition. a) Right Posterior STG: this region was commonly activated for "Spatial" vs "Non-spatial" scenes, independent of attention (yellow cluster in Figure 1c), b) Left Posterior STG: a similar pattern as in a), however, significance in this cluster was above the corrected threshold, c) Left PT and d) Left IFG: in these clusters the "Spatial" vs "Non-Spatial" was significant only during the attention to "Environment" condition (blue map in Figure 1c), e) Left middle STG, and f) Right middle STG: in these clusters the "Spatial" vs "Non-Spatial" was significant only during the attention to "Voice" condition (red map in Figure 1c).

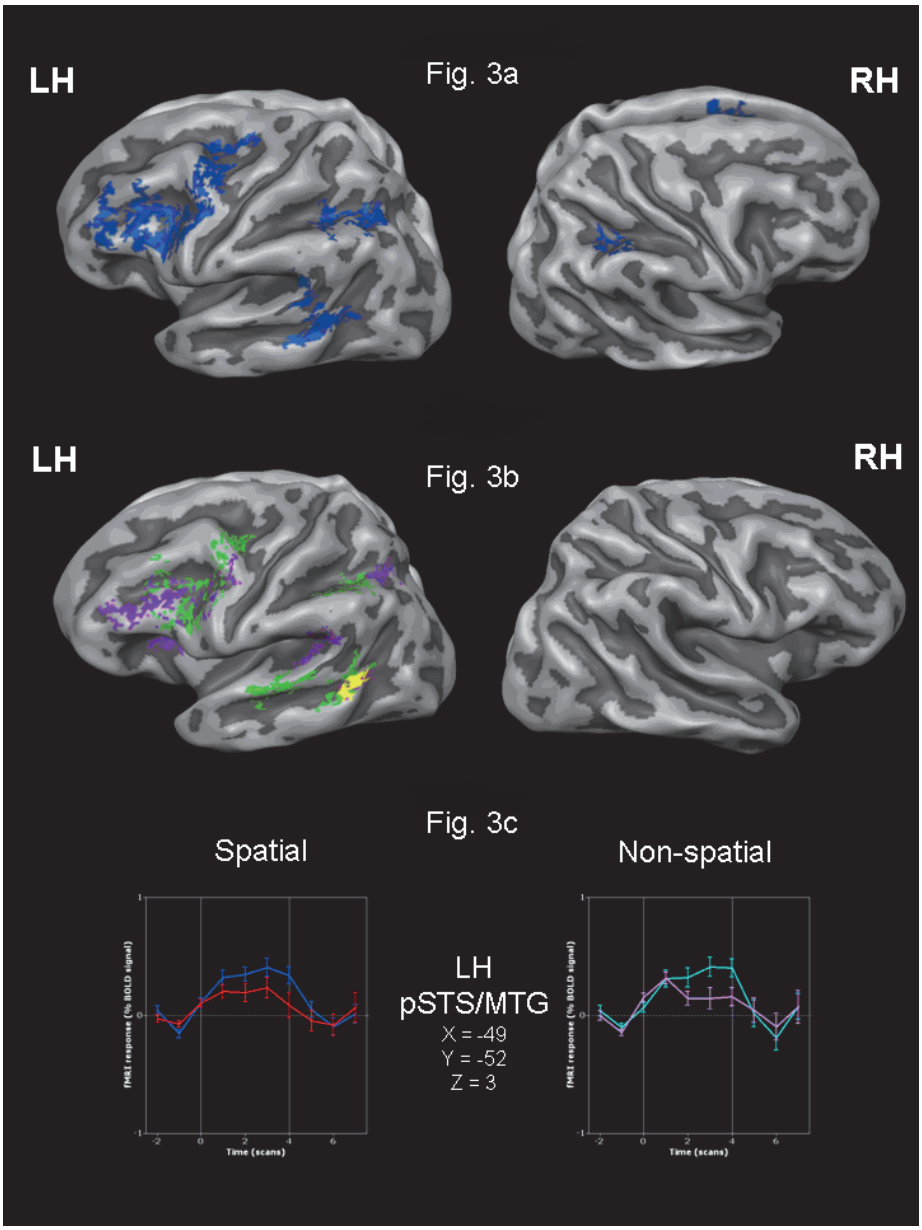

Figure 3: Results from Cortex-based Aligned Random Effect analysis using the General linear model. a) "Environment" versus "Voice" attention condition (merged spatial conditions), b) "Environment" versus "Voice" for either the "Spatial" scene condition (Purple) or the "Non-Spatial" condition (Green), c) Time courses related to fig. 3b (left posterior STS/MTG, yellow) where "Environment" (Blue and Turquoise lines) vs "Voice" (Red and Pink lines) scenes are compared during the "Spatial" (left column) or "Non-Spatial" (right column) attention condition: this area was commonly activated for "Environmental" vs "Voice" scenes, independent of spatial cues.

Discussion

In this study we investigated the cortical processes related to stream segregation of complex natural auditory scenes, with or without spatial acoustic information.

Comparing "Spatial" to "Non-Spatial" scenes resulted in a robust increase of regional activation in clusters of the right and - to a lesser extent - left posterior auditory cortex. This increased activation was present during both attention conditions (voice and environment). The anatomical location of these attention-independent activations corresponds to the posterior portion of the planum temporale, at a site which is compatible with area Tpt of the anatomical classification by (Galaburda and Sanides, 1980) (see also (Sweet et al., 2005)), area STA of the classification by (Rivier and Clarke, 1997) (see also (Wallace et al., 2002)) or area Te3 in the classification by (Morosan et al., 2005). These locations are also in agreement with previous functional neuroimaging studies that investigated sound localization and motion using simple sounds presented in isolation (Hart et al., 2004; Krumbholz et al., 2005a). Thus, in line with the general functional dichotomy between 'what' and 'where' auditory processing streams, our results confirm the indication that posterior auditory regions carry out the analysis of spatial cues in complex auditory scenes. Spatial processing in these areas seems to be automatic and scarcely influenced by attention, which may be particularly relevant for efficient localization of relevant and sudden sounds. It is worth noting that our experimental task did not explicitly require listeners to localize the sounds, which further highlights the obligatory nature of the observed effects.

Besides "automatic" sound localization, spatial acoustic cues from complex auditory scenes may also contribute to the processes of sound stream segregation and formation. Thus, the observed attention-dependent effects may reflect a second cortical processing mechanism, which may be devoted to integration of spatial cues with other spectral and temporal cues, with the goal of segregating and forming auditory streams. Such processing is expected in cortical locations specifically involved in the recognition of sounds. Also, attention is expected to have a relevant role in selecting and grouping the relevant sound object in the scene. Our results are highly consistent with this view. In regions of the middle portion of left (and right) STG and STS, we observed an effect of the spatial manipulation only when "voices" were attended to. These locations clearly resemble the so-called "voice sensitive" regions, as reported in previous studies (Belin et al., 2000). On the other end, we observed an effect of the spatial acoustic cues in regions of the left posterior Planum Temporale, only when attention was directed to background sounds. These regions have been associated in a previous study (Lewis et al., 2005) to processing of tool sounds, which in fact constitute a large subset of our background sounds.

Although consistent with previous studies, our interpretation is not univocal. In fact, in our scenes, voices were always located centrally in front of the listener, while the environmental sounds were peripheral and more variable. New studies should verify whether the observed attention-dependent effects reflect the different sensitivity of

auditory regions to distinct locations rather than the sensitivity to distinct sound categories, or relate them to each other.

Comparing the “Environment” vs “Voice” attention conditions resulted in a robust increase of activation mainly in the left temporal, frontal and bilateral parietal areas. On the contrary, the reverse contrast revealed no significant effect. These differences might be interpreted in the light of task difficulty. Attending to sounds on the periphery may require additional top-down signaling from frontal and parietal areas for overriding or counteracting the automatic allocation of attention to centrally-located or vocalized sounds.

Acknowledgements

Funding for the present research was contributed to E. Formisano (Vernieuwingsimpuls VIDI) from the Netherlands’ Organization for Scientific Research (NWO) and to H. Renvall from the Academy of Finland.

References

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., and Grady, C. L. (2001). "What" and "where" in the human auditory system. Proc Natl Acad Sci U S A 98, 12301-12306.

Altmann, C. F., Bledowski, C., Wibral, M., and Kaiser, J. (2007). Processing of location and pattern changes of natural sounds in the human auditory cortex. Neuroimage 35, 1192-1200.

Barrett, D. J., and Hall, D. A. (2006). Response preferences for "what" and "where" in human non-primary auditory cortex. Neuroimage 32, 968-977.

Baumgart, F., Gaschler-Markefski, B., Woldorff, M. G., Heinze, H. J., and Scheich, H. (1999). A movement-sensitive area in auditory cortex. Nature 400, 724-726.

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

Bregman, A. S. (1990). Auditory scene analysis (Cambridge, MA, MIT Press).

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., and Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn Reson Med 33, 636-647.

Galaburda, A., and Sanides, F. (1980). Cytoarchitectonic organization of the human auditory cortex. J Comp Neurol 190, 597-610.

Getzmann, S., and Lewald, J. (2010). Shared cortical systems for processing of horizontal and vertical sound motion. J Neurophysiol 103, 1896-1904.

Goebel, R., Esposito, F., and Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Hum Brain Mapp 27, 392-401.

Groh, J. M., Kelly, K. A., and Underhill, A. M. (2003). A monotonic code for sound azimuth in primate inferior colliculus. J Cogn Neurosci 15, 1217-1231.

Hart, H. C., Palmer, A. R., and Hall, D. A. (2004). Different areas of human non-primary auditory cortex are activated by sounds with spatial and nonspatial properties. Hum Brain Mapp 21, 178-190.

Krumbholz, K., Schonwiesner, M., Rubsamen, R., Zilles, K., Fink, G. R., and von Cramon, D. Y. (2005a). Hierarchical processing of sound location and motion in the human brainstem and planum temporale. Eur J Neurosci 21, 230-238.

Krumbholz, K., Schonwiesner, M., von Cramon, D. Y., Rubsamen, R., Shah, N. J., Zilles, K., and Fink, G. R. (2005b). Representation of interaural temporal information from left and right auditory space in the human planum temporale and inferior parietal lobe. Cereb Cortex 15, 317-324.

Lewis, J. W., Beauchamp, M. S., and DeYoe, E. A. (2000). A comparison of visual and auditory motion processing in human cerebral cortex. Cereb Cortex 10, 873-888.

Lewis, J. W., Brefczynski, J. A., Phinney, R. E., Janik, J. J., and DeYoe, E. A. (2005). Distinct cortical pathways for processing tool versus animal sounds. J Neurosci 25, 5148-5158.

Lomber, S. G., and Malhotra, S. (2008). Double dissociation of 'what' and 'where' processing in auditory cortex. Nat Neurosci 11, 609-616.

Miller, L. M., and Recanzone, G. H. (2009). Populations of auditory cortical neurons can accurately encode acoustic space across stimulus intensity. Proc Natl Acad Sci U S A 106, 5931-5935.

Morosan, P., Schleicher, A., Amunts, K., and Zilles, K. (2005). Multimodal architectonic mapping of human superior temporal gyrus. Anat Embryol (Berl) 210, 401-406.

Pavani, F., Macaluso, E., Warren, J. D., Driver, J., and Griffiths, T. D. (2002). A common cortical substrate activated by horizontal and vertical sound movement in the human brain. Curr Biol 12, 1584-1590.

Recanzone, G. H., Guard, D. C., Phan, M. L., and Su, T. K. (2000). Correlation between the activity of single auditory cortical neurons and sound-localization behavior in the macaque monkey. J Neurophysiol 83, 2723-2739.

Riecke, L., van Opstal, A. J., Goebel, R., and Formisano, E. (2007). Hearing illusory sounds in noise: sensory-perceptual transformations in primary auditory cortex. J Neurosci 27, 12684-12689.

Rivier, F., and Clarke, S. (1997). Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. Neuroimage 6, 288-304.

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., and Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat Neurosci 2, 1131-1136.

Smith, K. R., Hsieh, I. H., Saberi, K., and Hickok, G. (2010). Auditory spatial and object processing in the human planum temporale: no evidence for selectivity. J Cogn Neurosci 22, 632-639.

Sweet, R. A., Dorph-Petersen, K. A., and Lewis, D. A. (2005). Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. J Comp Neurol 491, 270-289.
Talairach, J., and Tournoux, P. (1988). Co-Planar Stereotactic Atlas of the Human Brain (Stuttgart, Thieme).
Tian, B., Reser, D., Durham, A., Kustov, A., and Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. Science 292, 290-293.
van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (2004). Integration of letters and speech sounds in the human brain. Neuron 43, 271-282.
Wallace, M. N., Johnston, P. W., and Palmer, A. R. (2002). Histochemical identification of cortical areas in the auditory region of the human brain. Exp Brain Res 143, 499-508.
Warren, J. D., and Griffiths, T. D. (2003). Distinct mechanisms for processing spatial sequences and pitch sequences in the human auditory brain. J Neurosci 23, 5799-5804.
Warren, J. D., Zielinski, B. A., Green, G. G., Rauschecker, J. P., and Griffiths, T. D. (2002). Perception of sound-source motion by the human brain. Neuron 34, 139-148.
Warren, J. E., Wise, R. J., and Warren, J. D. (2005). Sounds do-able: auditory-motor transformations and the posterior temporal plane. Trends Neurosci 28, 636-643.
Zatorre, R. J., Bouffard, M., Ahad, P., and Belin, P. (2002). Where is 'where' in the human auditory cortex? Nat Neurosci 5, 905-909.

CHAPTER 5

Brain-based un-mixing of vocal and instrumental streams during music listening

Corresponding publication:
Staeren, N.*, Valente, G.*, Formisano, E*, Brain-based un-mixing of vocal and instrumental streams during music listening (in preparation, *equal contribution)

Summary

Apart from being a common and engaging experience in our everyday life, listening to music well exemplifies the remarkable capabilities of the human auditory system to analyse sound mixtures. While listening to a song, one can distinctively perceive the singing voice and the various instruments despite the large spectral-temporal overlap between their sounds, a process known as stream segregation. In this study, we use music to reveal the mechanisms the human brain uses for processing multiple simultaneous auditory streams. During fMRI measurements, subjects were presented with two rock songs, which were played by the same group (voice [male singer], guitar, bass, and drum) but differed widely in terms of acoustic properties, melody, rhythm, spectro-temporal overlap of the streams. We show that a machine learning algorithm – trained with auditory cortical activation patterns elicited by one of the songs – can successfully decode the variations of acoustic energy in the singing voice and the other instruments from activation patterns elicited by the other song. For each of the sound sources (i.e. the voice and the instruments), informative patterns comprised distinct – yet overlapping - networks of superior temporal regions.

These findings indicate that the brain processing of a complex sound mixture (such as a song) involves the formation of neural representations of each contributing source. The successful decoding of sound sources across mixtures that differed along multiple acoustic dimensions suggest that these auditory cortical representations are perceptual rather acoustic.

Introduction

Listening to music is an everyday experience that well exemplifies the level of computational complexity our auditory system continuously faces. When listening to a song, we can distinctively perceive the singing voice and various instruments, despite the large spectral and temporal overlap of their sounds. Under this perspective, music can be regarded as a rich and natural stimulus for studying auditory scene analysis (ASA). ASA refers to the processes required of the auditory system to recover descriptions of individual sound sources ('auditory objects' or 'auditory streams') from mixtures of simultaneous sounds (Bregman, 1990). Although studied extensively in psychophysics (Bregman, 1990; Ciocca, 2008), little is known about the neural mechanisms underlying ASA. Several neuroimaging investigations in humans or invasive recordings in animals have examined the neural correlates of ASA using very elementary auditory scenes (e.g. tones in noise or alternating tone streams). Results from these studies suggest a relevant role of the primary auditory areas in the formation of simple auditory objects; furthermore they have put forward a number of coding mechanisms that the brain may use for solving these elementary ASA problems (Eggermont, 2001; Elhilali et al., 2009). The simplicity of the scenes examined, however, does not allow generalizing the obtained results to more complex and realistic scenes. Processing and segregating realistic auditory scenes certainly involve additional cortical representations and computations which have not been identified so far.

In the present study we combine music, functional magnetic resonance imaging (fMRI) and advanced computational methods to address the neural foundations of ASA. Numerous neuroimaging studies have already examined the neuronal basis of different aspects of music perception (see (Levitin and Tirovolas, 2009; Peretz and Zatorre, 2005) for reviews). Most of these studies, however, report the neural correlates of music listening at the level of activated brain regions (Levitin and Tirovolas, 2009; Peretz et al., 2009; Zatorre et al., 2007) or have focused on specific aspects of music processing, such as expectation violation (e.g. (McDermott and Oxenham, 2008)). Our goal, however, largely differs from these studies as it considers music as a means to unravel the neural make up of auditory objects or auditory streams. We define an auditory object or auditory stream as the perceptual – rather than physical – description of a sound, which is invariant to acoustic variations and to the background noise (Griffiths and Warren, 2004). Formation of these invariant representations is a crucial computational step in the analysis of any auditory scene. Here we exploit recent advances in fMRI data analysis (Formisano et al., 2008a; Formisano et al., 2008b; Valente et al., 2011) to investigate – during music listening - the formation of cortical representations of individual instruments and voice which are invariant to the melody and to the specific combination of the instruments being played. During fMRI measurements subjects listened to two rock songs played by different combinations of three instruments (guitar, bass and drums) and a singing voice, which have been recorded separately and mixed together in a professional setting (see Recording of stimuli). Using multivariate regression (Gaussian

processes), we first estimate a "brain signature" of each stream (i.e. the three instruments and the voice) by using fMRI data relative to one of the songs. This brain signature is obtained as the distributed brain representations ('predictive maps') associated with the set of continuous predictors corresponding to the root mean square (RMS) power of each instrument/voice. Then the robustness of this distributed representation is tested by looking at the capability of the brain-based machine learning algorithm to predict – in the second song - the individual streams. In this phase, information on the song is not given to the algorithm, but the individual tracks of the instruments/voice (RMS) are blindly predicted based on the fMRI data and on the predictive maps obtained in the learning phase. Successful learning is assessed by comparing the generated predictions with the original RMS profiles of the instruments and voice (correlation). Based on recent results with non-musical sounds presented in isolation (Formisano et al., 2008a; Staeren et al., 2009), we expected the invariant representations of the instruments' energy measures (RMS) to be comprised within the auditory regions of the superior temporal cortex. This would indicate that beyond the physical representation of sounds, these auditory regions entail an abstract level of stream representation.

Methods

Experimental setup

A scheme of the experimental protocol is illustrated in Figure 1. Subjects listened to two songs of 5 minutes, which were presented nine times each, with different combinations of instruments and voice: the three instruments and voice in solo, four mixtures of three instruments and/or voice, and the complete song. Each song presentation was interleaved with a short period of silence (30 sec) before the next song was presented. The order of the runs was balanced across the subjects. Songs were presented binaurally using mono playback. Subjects were instructed to attentively listen to the music as if it would be played on the radio. Subjects were unfamiliar with the music.

Recording of stimuli

The songs were two new rock compositions. This type of music was chosen because it is relatively 'simple' to record and is easily accessible by people with different musical backgrounds. Instrument tracks were recorded (48 KHz/32 bit) with aid of a professional sound engineer. One of the authors (NS) co-created the music, edited and mastered each instrument (guitar, bass, drum) and voice separately using a Protools LE setup (Digidesign, Daly City, CA, USA) (see Appendix 1 for detailed information on the recording and sound editing procedures). By using own recordings, we were able to quantify the contribution of each instrument to the overall physical sound signal, as well as to create songs consisting of the different combination of instruments and voice. No master recording (with separate track per instrument) is usually available for commercial and copyrighted music. During the

measurements, the stimuli were delivered binaurally via MR compatible headphones (Com-
(Commander XG, Resonance Technology, Northridge, CA) at a comfortable listening level.

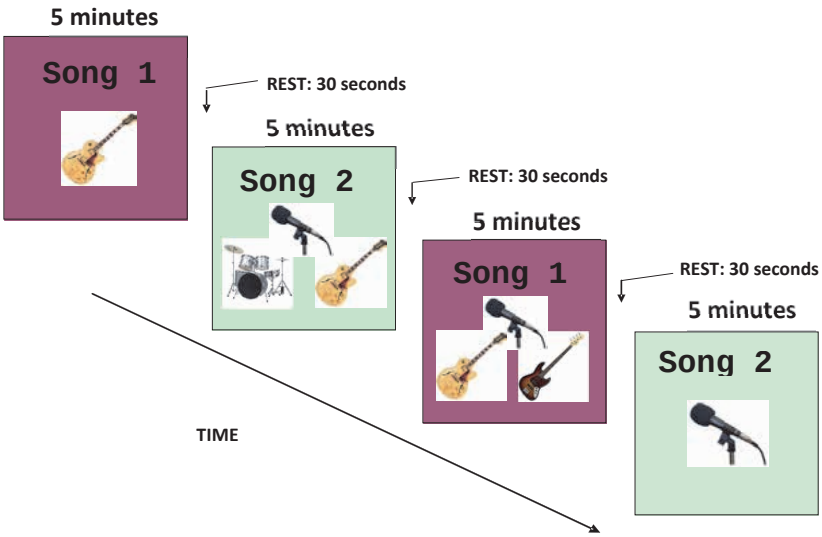

Figure 1: Experimental design of a run. Each song was presented nine times, with different combinations of instruments. The two songs are interleaved, with a 30 second rest period in between.

Subjects

Four subjects that gave their informed consent, (mean ± SEM age 27 ± 4 yrs; 2 females; 2 right-handed) participated in the study. The subjects were graduate university students who were paid for their participation. Subjects had no history of hearing or neurological impairments, and were naïve to the experimental setup and music. The study received a prior approval by the Ethical Committee of the Faculty of Psychology and Neuroscience, University of Maastricht.

fMRI measurements

Brain imaging was performed with a 3 Tesla Siemens Allegra (head setup) at the Maastricht Brain Imaging Center. In each subject, two runs of 850, and two runs of 682 volumes were acquired with a T2-weighted gradient-echo planar imaging (EPI) sequence (TR = 2000 ms, voxel size = 2,5 × 2,5 × 3 mm^3, TE = 30 ms, FOV 256 × 256; matrix size 96 × 96, 31 slices covering the cortex). Four runs consisted of four or five songs (different instrument combinations) and lasted approximately 22 or 28 min. Anatomical images were obtained using a 1 × 1 × 1 mm^3 resolution T1-weighted MPRAGE sequence between the second and third functional run.

fMRI Data Analysis: pre-processing, univariate and multivariate statistics

Functional and anatomical images were first analyzed with BrainVoyager QX (Brain Innovation, Maastricht, The Netherlands). Preprocessing consisted of slice scan-time correction (using sinc interpolation), linear trend removal, temporal high-pass filtering to remove nonlinear drifts of seven or less cycles per time course, and 3-dimensional motion correction. Temporal low pass filtering was performed using a Gaussian kernel with FWHM of two data points. Functional slices were co-registered to the anatomical data, and both data were normalized to Talairach space (Talairach and Tournoux, 1988).

FMRI time series were modeled using multivariate regression (Gaussian Process Regression (GPR)) with a linear covariance function having one hyperparameter. (See (Rasmussen and Williams, 2006)) for more details on GPR). In order to relate the tracks to fMRI activity we constructed a predictor for each instrument and voice by computing - separately for each individual track - the average Root Mean Square (RMS) power within a volume acquisition. This RMS time course was then convolved with an estimate of the hemodynamic response function (HRF), (See Figure 2). These predictors were used to train four GPR models (one for each instrument and voice). Because a linear covariance function was used, this training also resulted in four distinct predictive maps, i.e. maps coding the relative importance of voxels in predicting new data. The generalization performance of these models/maps was then tested through the prediction of the corresponding RMS profile/predictors in the song not used during the training. Successful learning was assessed by comparing the generated predictions with the original RMS profiles of the instruments and voice (correlation: value and significance). Significance (in the form of a p-value) was assessed at the group level (random effects, 4 subjects). Correlation values were transformed with Fisher transform, and a t-test was used to compare the obtained correlations with the theoretical chance level of r= 0.

Results

The multivariate algorithm was able to learn the relationship between the instruments' RMS profile and corresponding brain activation patterns. This was assessed by training the algorithm in one song and testing the accuracy of decoding the same instruments from brain activation in the other different song.

Figure 3a and 3b illustrates the original (black line) and predicted RMS profile (blue line) for voice, guitar, bass, and drums, trained on the first song and tested on the second song (Figure 3a) and vice versa (Figure 3b). In the left column, the RMS profile refers to listening to an individual instrument (“Instrument alone”). In the right column, the RMS profile of an instrument was predicted using brain data during natural music listening (“Full song mix”). The RMS profile predictions have been averaged across the four subjects, with the gray area containing values within one standard error.

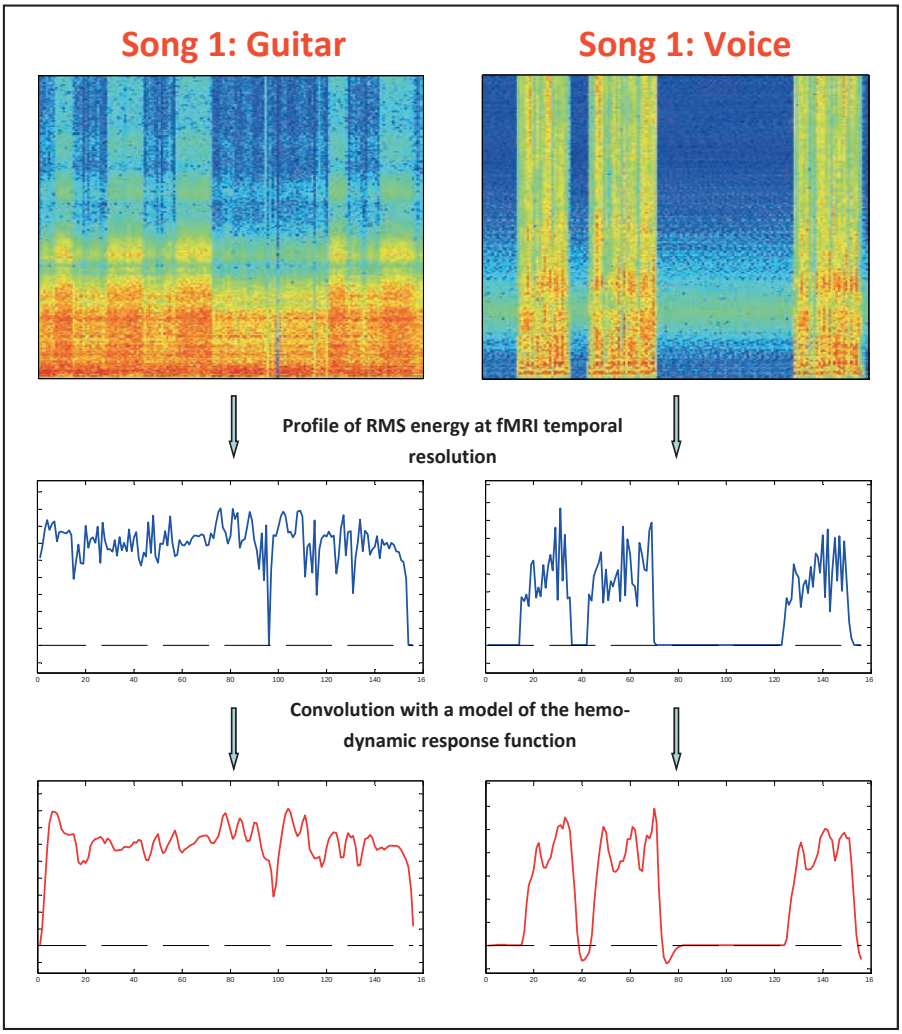

Figure 2: Construction of predictors for guitar and voice. For each individual track, we computed the average Root Mean Square (RMS) power within a volume acquisition. This RMS time course was then convolved with an estimate of the hemodynamic response function (HRF),

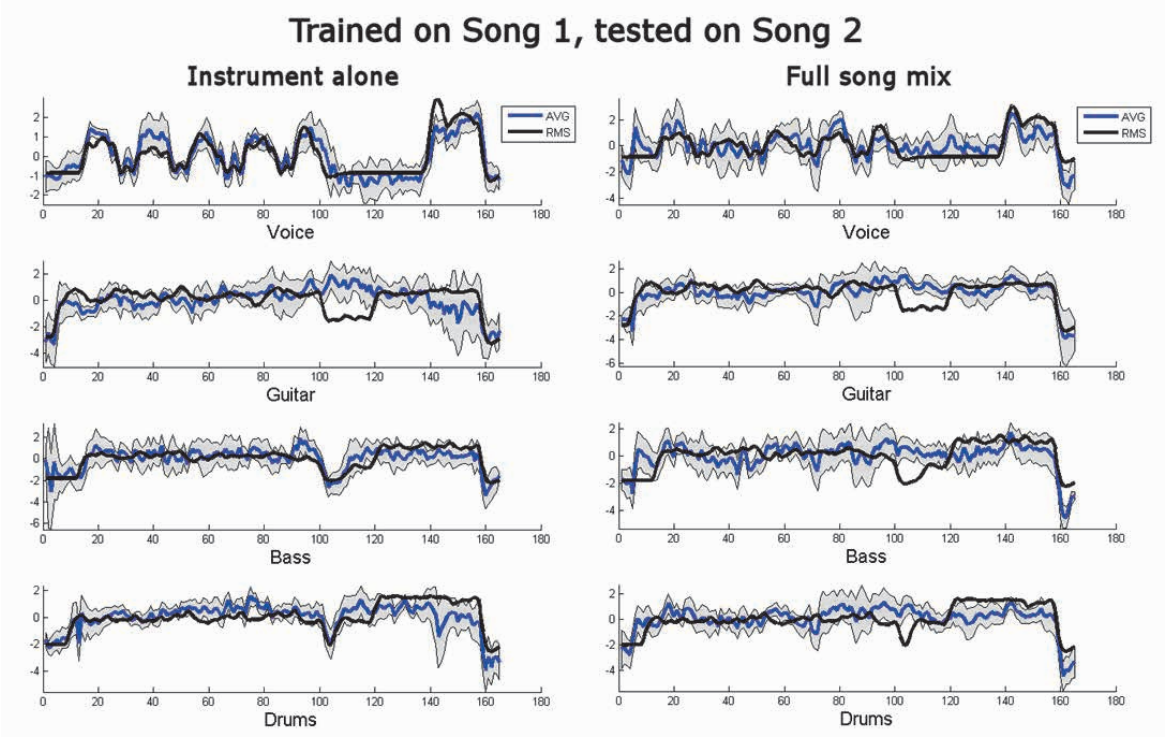

Figure 3a: After learning the relationship between the instruments' RMS values and brain activation of one song, the multivariate algorithm decoded the Instrument sources (RMS) from brain activation of the 'unknown' song within different mixtures, during individual instrument listening (left column) and during natural music listening ("Full song mix", right column). The original (black line) and brain-based prediction (blue line, average of 4 subjects, with the gray area containing values within one standard error.) are illustrated for voice, guitar, bass, and drums, trained on song one and tested on song two.

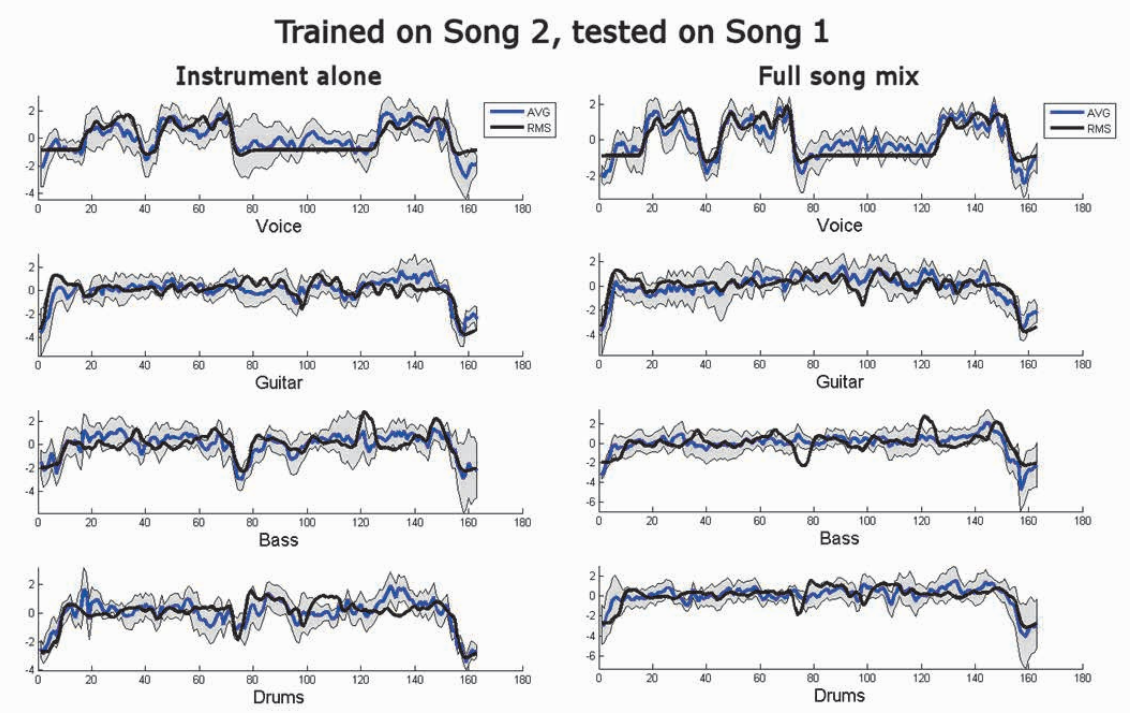

Figure 3b: After learning the relationship between the instruments' RMS values and brain activation of one song, the multivariate algorithm decoded the Instrument sources (RMS) from brain activation of the 'unknown' song within different mixtures, during individual instrument listening (left column) and during natural music listening ("Full song mix", right column). The original (black line) and brain-based prediction (blue line, average of 4 subjects, with the gray area containing values within one standard error.) are illustrated for voice, guitar, bass, and drums, trained on song two and tested on song one.

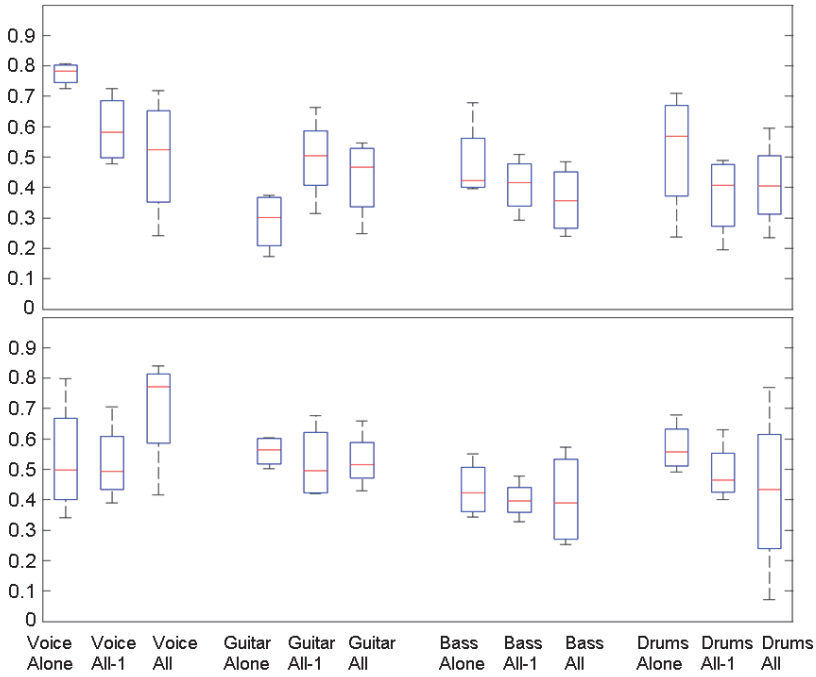

Figure 4: For each instrument and scene complexity (one, three (average) or four instruments), the average correlation values are displayed between predicted and original RMS profile when training on song 2 and testing on song 1 (upper plot), and vice versa (lower plot).

Figure 4 shows the box plots of the correlations between predicted and original RMS profiles (between songs) relative to different mixtures levels of the test song (solo (Alone), 3 instruments (average extraction result of all 3 instrument combinations (All-1)), and the full song (4 instruments (All)). As there were more combinations of 3 instruments, the correlations for the three combinations were first transformed with Fisher transformation and then averaged and transformed back with inverse Fisher transformation. The Median values of correlations (over subjects) when training on song 1 and testing on song 2 (Figure 4, upper plot) and vice versa (Figure 4, lower plot), within different mixture levels (instrument alone, 3 instruments, complete arrangement) are reported in Table 1 together with their significance value.

In all cases, the correlation values were above chance, which indicates that the prediction of the instruments' RMS values from brain activation of a different song was robust to substantial overlap and variations of the acoustic sources.

Table 1: The Median values of correlations (over 4 subjects) when training on song 1 and testing on song 2 (Upper table) and vice versa (lower table), within different mixture levels (instrument alone (Alone), *3 instruments (All-1), complete arrangement (All))*

	Alone		All -1		All	
	correlation	p-value	correlation	p-value	correlation	p-value
Trained on song 1 and tested on song 2						
Voice	0.7829	0.0002	0.5821	0.0047	0.5244	0.0249
Guitar	0.3016	0.0108	0.5039	0.0102	0.4675	0.0097
Bass	0.3557	0.0099	0.4232	0.0117	0.4157	0.0043
Drums	0.5680	0.0222	0.4062	0.0136	0.4041	0.0165
Training on song 2 and tested on song 1						
Voice	0.4982	0.0287	0.4942	0.0096	0.7717	0.0118
Guitar	0.5647	0.0004	0.4952	0.0069	0.5161	0.0033
Bass	0.3894	0.0196	0.4225	0.0040	0.3959	0.0014
Drums	0.5576	0.0019	0.4636	0.0043	0.4337	0.0812

To identify informative brain sites for the RMS-based stream segregation, we estimated a predictive map for each instrument (Figure 5). These maps represent the voxels which contributed most to the extraction of an individual instrument tracks (RMS profile) from unknown song data, averaged for the different levels of mixture extraction. The areas of the different streams were mostly localized in superior temporal cortex bilaterally and were partly overlapping. Figure 5a illustrates these maps for 'voice' (blue) and for 'guitar' (red). At the group level, the richest sources of information for predicting the 'voice' stream were located at the left posterior/middle planum polare (PP), anterior Heschl's gyrus (HG), planum temporale (PT) and middle superior temporal gurys/superior temporal sulcus (STG/STS) in the left hemisphere, and at the middle PP, anterolateral HG, Heschl's sulcus (HS) and middle STG/STS in the right hemisphere. This predictive map for 'voice' included regions located in middle and anterior STG/STS and is in spatial agreement with studies investigating the processing of voices (Belin et al., 2000; Formisano et al., 2008a). The guitar was best predicted using voxels from the anterior/middle PP, anterior HG in the left hemisphere, and at the middle PP, lateral HG, HS and middle STG in the right hemisphere. Figure 5b illustrates the predictive maps for 'bass' (green) and 'drums' (pink). The 'bass' was best predicted using voxels located at the middle PP, anterior HG, HS and anterior PT in the left hemisphere, and middle PP, HG, HS and anterior PT in the right hemisphere. 'Drums' were best predicted using voxels located at the middle PP, HG, HS and anterior PT in the left hemisphere, and middle PP, HG, HS, middle STG and PT in the right hemisphere.

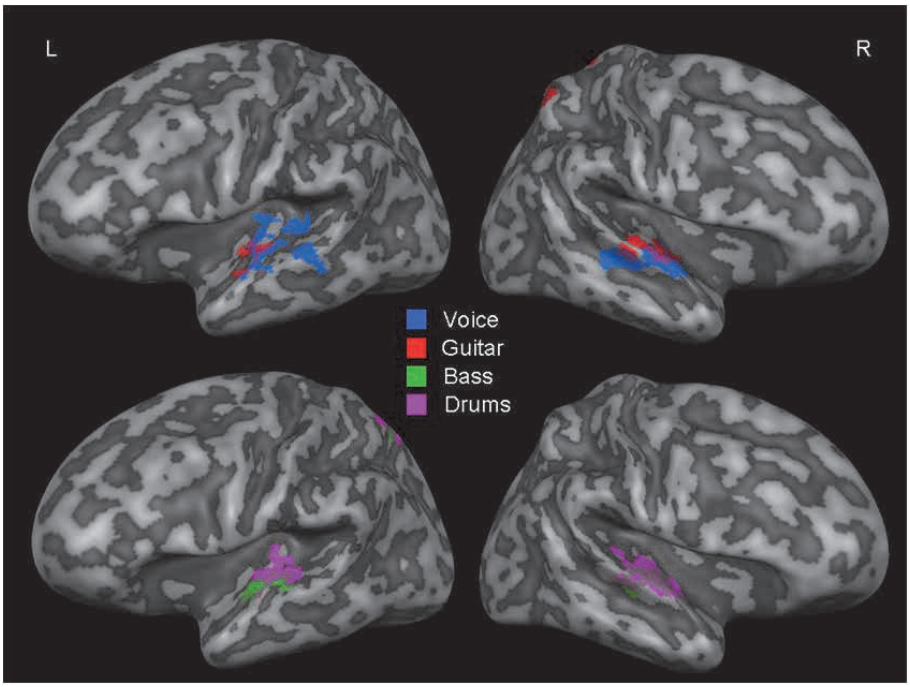

Figure 5: Predictive maps for 'voice' (blue), 'guitar' (red), 'bass' (green) and 'drums' (pink).

Discussion

The results of this study demonstrated that it is possible to un-mix the fMRI responses measured while subjects listen to complex mixtures of sound streams (i.e. rock songs) into separated response patterns, each one encoding for an individual stream (i.e. voice and instruments). With multivariate regression, we examined the relation between a continuous descriptor (RMS energy) of each stream and whole-brain fMRI responses patterns. We revealed that the simultaneous representation of multiple sources during auditory stream segregation is encoded in distributed and partially overlapping cortical networks in the superior temporal cortex. Our results suggest that these source representations are perceptual rather than acoustic. In fact, the successful prediction of instrumental track/voices across songs implies that the estimated patterns convey information on the sources which is robust to large acoustic variations of the sources and to the presence/variation of all other intervening sources. In other words, it implies that the multivariate modeling has "filtered" out the effects on the brain responses of the background and that the resulting representations contain information on more abstract dimensions of the source beyond simple acoustics.

Our approach - which does not require a strict control of the acoustic properties of the streams - enabled us studying ASA with a naturalistic and complex stimulus such as music

and under very realistic listening conditions. Such an experimental design differs largely from previous studies on the neural basis of ASA. In most cases, previous ASA studies considered elementary auditory scenes and gestalt-based streaming paradigms, e.g. with alternating tones to create auditory streams (e.g. (Fishman et al., 2001; Micheyl et al., 2007; Petkov et al., 2007)). Besides frequency and time separation, however, natural auditory scenes include many additional cues (such as timbre) that the auditory system can exploit for segregating the mixture and creating a perceptual stream. It is thus with these complex scenes that the cortical areas and mechanisms relevant for ASA may be optimally engaged and studied (Snyder and Alain, 2007).

Despite the substantially different methods employed, it is useful to compare our predictive maps with conventional statistical maps obtained in previous functional neuroimaging studies. In particular, our predictive map for 'voice' presented peaks localized bilaterally in "early" auditory regions (HG) as well as in "voice sensitive" regions of the middle and anterior STG/STS (Belin et al., 2000). These results are in agreement with our previous studies suggesting that response patterns in a set of early as well as higher level auditory areas encode voice and speaker identity, independently of content (Formisano et al., 2008a). The interpretation of the predictive maps for the other instruments remains more difficult as previous studies useful for the comparison are scarce. Informative locations for "bass" and "drum" maps appear to occupy mostly regions extending in the posterior HG and PT, which is consistent with a recent study suggesting the involvement of the (left) planum temporale in the processing of rhythmic structure (Herdener et al., 2012). Conversely, informative locations for 'guitar' (red) are more medial and anterior, which is consistent with the reported involvement of these auditory regions for the processing of FM sweeps. Furthermore, this dissociation is consistent with a recent model (Santoro et al., submitted) suggesting that regions posterior and anterior to the HG differ in terms of their high temporal resolution (posterior) and spectral resolution (anterior), which may be related respectively to the representation of rhythm and melody.

Finally, it is worth noting that in the present study we used RMS, an estimate of energy, as a continuous descriptor of each instrument within a mix. However, to study other aspects of acoustic, perceptual or emotional processing of music, similar analyses can be performed with different descriptors of the individual streams or of the musical piece as a whole.

Acknowledgements

We would like to thank Dries D'Hondt (Sound engineer, 3S, Leuven, Belgium) for recording and for professional assistance during the recording, Don Demuynck, Jan Jeurissen and Jonas Simons for participating as musicians and co-composers. Funding for the present research was contributed to E. Formisano (Vernieuwingsimpuls VIDI) from the Netherlands' Organization for Scientific Research (NWO).

References

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

Bregman, A. S. (1990). Auditory scene analysis (Cambridge, MA, MIT Press).

Ciocca, V. (2008). The auditory organization of complex sounds. Front Biosci *13*, 148-169.

Eggermont, J. J. (2001). Between sound and perception: reviewing the search for a neural code. Hear Res *157*, 1-42.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., and Shamma, S. A. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron *61*, 317-329.

Fishman, Y. I., Reser, D. H., Arezzo, J. C., and Steinschneider, M. (2001). Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hear Res *151*, 167-187.

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008a). "Who" is saying "what"? Brain-based decoding of human voice and speech. Science *322*, 970-973.

Formisano, E., De Martino, F., and Valente, G. (2008b). Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Magn Reson Imaging *26*, 921-934.

Griffiths, T. D., and Warren, J. D. (2004). What is an auditory object? Nat Rev Neurosci *5*, 887-892.

Herdener, M., Humbel, T., Esposito, F., Habermeyer, B., Cattapan-Ludewig, K., and Seifritz, E. (2012). Jazz Drummers Recruit Language-Specific Areas for the Processing of Rhythmic Structure. Cereb Cortex.

Levitin, D. J., and Tirovolas, A. K. (2009). Current advances in the cognitive neuroscience of music. Ann N Y Acad Sci *1156*, 211-231.

McDermott, J. H., and Oxenham, A. J. (2008). Music perception, pitch, and the auditory system. Curr Opin Neurobiol *18*, 452-463.

Micheyl, C., Carlyon, R. P., Gutschalk, A., Melcher, J. R., Oxenham, A. J., Rauschecker, J. P., Tian, B., and Courtenay Wilson, E. (2007). The role of auditory cortex in the formation of auditory streams. Hear Res *229*, 116-131.

Peretz, I., Gosselin, N., Belin, P., Zatorre, R. J., Plailly, J., and Tillmann, B. (2009). Music lexical networks: the cortical organization of music recognition. Ann N Y Acad Sci *1169*, 256-265.

Peretz, I., and Zatorre, R. J. (2005). Brain organization for music processing. Annu Rev Psychol *56*, 89-114.

Petkov, C. I., O'Connor, K. N., and Sutter, M. L. (2007). Encoding of illusory continuity in primary auditory cortex. Neuron *54*, 153-165.

Rasmussen, C. E., and Williams, C. K. I. (2006). Gaussian Processes for Machine Learning, The MIT Press).

Santoro, R. Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E. and Formisano, E. (Submitted) Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex.

Snyder, J. S., and Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. Psychol Bull *133*, 780-799.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., and Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. Curr Biol *19*, 498-502.

Talairach, J., and Tournoux, P. (1988). Co-Planar Stereotactic Atlas of the Human Brain (Stuttgart, Thieme).

Valente, G., De Martino, F., Esposito, F., Goebel, R., and Formisano, E. (2011). Predicting subject-driven actions and sensory experience in a virtual world with relevance vector machine regression of fMRI data. Neuroimage *56*, 651-661.

Zatorre, R. J., Chen, J. L., and Penhune, V. B. (2007). When the brain plays music: auditory-motor interactions in music perception and production. Nat Rev Neurosci *8*, 547-558.

Appendix 1

The songs were recorded in a sound isolated studio room. The following standard recording process was followed: All musicians simultaneously played the complete song, which was recorded as the reference track. Then individual musicians played their individual instrument track, recorded on separate channels, while listening to the reference track through their headphones. All microphones were connected to a Soundtracs Solo Midi Production Console (DiGiCo UK Ltd, Surrey, UK) which was also used for preamplification. From here 16 channels were connected to a Protools LE setup (Digidesign, Daly City, CA, USA) which recorded the sound at 48Khz/32bit on a personal computer. Recordings were performed and monitored by a professional sound engineer (Dries D'Hondt, 3S, Leuven, Belgium). An Epiphone Sheraton electric guitar (Gibson Guitar Corp, Nashville, TN, USA) amplified with a Mesa Boogie Combo (Mesa Boogie Ltd, Petaluma, CA, USA) was recorded on three tracks: 1) Analog Amplifier line out, 2) One Shure SM57 (Shure Electronics, Niles, IL, USA) microphone placed against the combo speaker 3) One Sennheiser MD 421 (Sennheiser Electronic GmbH & Co. KG, Wedemark, Germany) placed against the combo speaker. The bass guitar was recorded using a Fender Jazz Bass USA electric bass guitar (Fender, Scottsdale, AZ, USA), amplified with a SWR Basic Black edition (SWR Sound Corporation, Scottsdale, AZ, USA). A Sennheiser MD 421 microphone was positioned against the amplifier's speaker for recording the first bass track. The second bass track was recorded using a BSS AR116 direct inject box (BSS Audio, Sandy, UT, USA) which was connected to the D.I. output of the amplifier. Vocals were recorded using a Røde NT2 microphone (Røde Microphones, Silverwater, Australia). A Yamaha Stage Custom drumkit (Yamaha Corporation, Hamamatsu, Japan) and Sabian cymbals (Sabian Inc, Marshfield, MA, USA) XS Medium tin crash, XS crash ride, Pro Sonix China, AA rock ride, AA splash and AA rock hi-hat were recorded using the following setup: Bass drum (Audix D6 microphone (Audix Microphones, Wilsonville, OR, USA), snare drum (Shure SM57), toms 1,2 and 3 (floortom) (3 Audiotechnica Pro 35 microphones (Audio-Technica, Tokyo, Japan)). Cymbals were recorded using two Oktava MK 012 (OAO Oktava, Tula, Russia) microphones. One Røde NT2 was used for the room recording of the drumkit. Minimal mastering was performed in Pro Tools using a Pro Tools Mbox system (Digidesign, Daly City, CA, USA) to preserve the recording's original spectral-temporal pattern and to avoid unpredictable spectral-temporal modification of the mastering process which could hinder computational stream segregation. Individual tracks were then imported in Adobe Audition (Adobe Systems Inc., Mountain View, CA, USA) at 48000Hz/32bit, which was used to combine the tracks into the required instrument combinations. These were exported as 44000Hz/16bit mono mixes for use in Presentation 9.3™ (Neurobehavioral Systems, Inc., Albany, CA, USA).

Summary

This thesis describes functional neuroimaging (fMRI and MEG) research designed to study the relationship between human brain activity and the perception of natural sounds. Many studies in the field of auditory neuroscience use synthetic sounds to investigate auditory perception. Synthetic sounds allow researchers a great level of control over the physical parameters of the stimulus, making them more suitable for understanding the neural processing of basic acoustic features. The four studies presented here take the complementary perspective of using natural sounds to explore the brain mechanisms for sound categorization and auditory stream segregation under realistic and ecologically valid conditions. Also in terms of analysis methods employed, the described studies present relevant differences with previous research. So far, the vast majority of functional neuroimaging studies investigated sound categorization using subtraction-based experimental paradigms and conventional univariate (voxel-by-voxel) statistics. These paradigms and statistical methods are inherently bound to produce results in terms of 'specialization' or 'selectivity' for a certain stimulus attribute or category, as they can only detect localized surplus of hemodynamic activity for one condition compared to another, possibly ignoring potential information which could be represented in non-maximal responses. For this reason, two of the presented studies (chapters 2 and 5) make use of multivariate analysis methods. These methods allow modeling the functional relation between spatial patterns of brain activity and stimulus categories (chapter 2, multivariate classification) or continuous variations in the stimulus (chapter 5, multivariate regression). Beyond looking at subtractive contrasts that differentiate conditions, with these methods the similarity among response patterns under changing stimulus conditions can be tested. Such possibility is pivotal to address relevant questions on the neural underpinnings of auditory perception, such as the invariance of categorical neural representations to changes of low level acoustic properties (chapter 2) or to changes of the acoustic background (chapter 5).

The first part of the thesis (**chapter 2** and **3**) investigates the neural mechanisms of sound recognition using natural sounds presented in isolation and functional neuroimaging at high spatial resolution (fMRI, chapter 2) and high temporal resolution (MEG, chapter 3). In **Chapter 2**, sounds from four categories (cats, female singers, acoustic guitars, and tones) were recorded, carefully matched for their time-varying spectral characteristics and presented to subjects at three different pitch levels. Univariate contrasts between categories

did not lead to statistically significant effects, suggesting that the control on the acoustic sound properties largely reduces the differences of regional BOLD responses, which are often observed when comparing different sound categories. Sound category information - not detectable using voxel-by-voxel analysis - could be instead detected and mapped with multivoxel pattern analyses. Encoding of sound 'category' independent of pitch was spatially distributed over a large expanse of the bilateral supratemporal cortices, whereas a more localized pattern was observed for encoding of 'pitch' laterally to primary auditory areas. These results suggest that the conventional regional effects (found e.g. in "voice mapping" measurements) mostly reflect the processing of multiple acoustic features. Conversely, more abstract 'categorical' representations of natural sounds may emerge from the joint encoding of information occurring not only in this small set of higher-level selective areas but also in auditory areas conventionally associated with lower-level auditory processing.

The study in **Chapter 3** exploits the high temporal resolution of MEG measurements to investigate the time-course of sound categorization in the presence of minimal or no acoustic differences among the incoming stimuli. Female voices and cat sounds from chapter 2 were further manipulated and filtered so they matched in most of their acoustic properties. A "category priming" paradigm was used that allowed to examine auditory cortical processing of two categories beyond the physical make-up of the stimuli, using MEG. During the measurements, a category context was established, followed by a probe sound that was congruent, incongruent, or ambiguous to this context. The results show that MEG responses to incongruent sounds were stronger than responses to congruent sounds at ~250 ms in the right temporoparietal cortex, regardless of the sound category. Furthermore, probe sounds that could not be unambiguously attributed to any of the two categories ("cat" or "voice") evoked stronger responses after the voice than cat context at 200–250 ms, suggesting a stronger contextual effect for human voices.

Taken together, the findings of these two studies indicate that distributed neuronal populations within the human auditory areas entail categorical representations of sounds, beyond their physical properties. Categorical templates for human and animal vocalizations seem to be established at ~250 ms from stimulus onset.

Chapters 4 and **5** form the second part of the thesis, which studies the neural basis of 'auditory scene analysis' with fMRI. Auditory scene analysis refers to the processes required for deriving descriptions of individual sound sources ('auditory objects' or 'auditory streams') from mixtures of simultaneous sounds. Because natural environments typically involve multiple sound sources, auditory scene analysis represents a crucial aspect of hearing, which lies at the heart of the ability to select and respond to relevant acoustic stimuli even when these are masked by competing sound sources or background noise.

Chapter 4 focuses on the cortical processing of spatial cues during listening to natural auditory scenes. Using the technique of binaural recording and in-ear microphones, realistic auditory scenes were recorded that contained two concurrent sounds, a human voice

centrally located in front of the listener (foreground), and an environmental sound located at different locations at the background. During fMRI measurements subjects were instructed to attend one of the sound sources ("*Voice*" vs "*Environment*"), under two distinct playback conditions: 1) Stereo playback which preserves the spatial acoustic information of the original recordings ("*Spatial*") or 2) Mono playback, which removes spatial information ("*Non*-spatial"). The statistical analyses showed that processing of the spatial cues - independently of the attention condition - corresponded with significantly increased brain activation at the bilateral posterior superior temporal areas. These regions are known for processing spatial and sound motion information (auditory "where" stream). However, significant activation differences in the *Spatial* vs *Non-spatial* comparison were observed that depended on the attention target. When listeners attended to environmental background sounds, we found significant differences in left planum temporale and left inferior frontal gyrus. Conversely, when listeners attended to vocal sounds, significant activation differences were found in bilateral clusters of middle superior temporal gyrus and sulcus, which overlap with the so called "voice sensitive regions". These attention-dependent effects suggest that – in order to segregate an auditory source from a sound mixture - spatial cues are integrated with other relevant spectral and temporal cues in the same cortical locations involved in the recognition of sounds presented in silence.

In the study described in **Chapter 5**, music is used to reveal the mechanisms the human brain uses for processing multiple simultaneous auditory streams. In contrast to chapter 4, where scenes included combinations of short auditory events, the auditory scenes in this chapter are mixtures of sound streams that are prolonged over time. During fMRI measurements, subjects were presented with two rock songs, which were played by the same group (voice [male singer], guitar, bass, and drum) but differed widely in terms of acoustic properties, melody, rhythm, spectro-temporal overlap of the streams. Results showed that a machine learning algorithm of multivariate regression – trained with auditory cortical activation patterns elicited by one of the songs – could successfully decode the variations of acoustic energy in the singing voice and the other instruments from activation patterns elicited by the other song. For each of the sound sources (i.e. the voice and the instruments), informative patterns comprised distinct – yet overlapping - networks of superior temporal regions. These findings indicate that the brain processing of a complex sound mixture (such as a song) involves the formation of neural representations of each contributing source. The successful decoding of each sound source across mixtures that differed along multiple acoustic dimensions suggest that these auditory cortical representations are perceptual rather acoustic. The highest decoding accuracy obtained for the vocal stream, which is most likely the "default" target of attention during music listening, suggests that the attended stream (foreground) is enhanced with respect to the other streams (background).

In sum, the results of chapter 4 and 5 indicate that neural sound representations in auditory networks in the superior temporal cortex are crucial for both bottom-up processing of spectral and temporal relations of the acoustic scene elements and top-down processes of attentive selection and enhancement of the relevant sounds. The range of methods and experimental paradigms introduced in this thesis pave the way for further studying the nature and computational properties of these representations, while probing the brain under the ecologically and behaviorally valid conditions of “real life” listening.

Acknowledgements

This thesis could not have been completed without the help of the following people. Elia, thank you for giving me the opportunity to work with you throughout the years. Your drive and passion for this work made you a unique supervisor that was able to deal with scientific as well as human difficulties during the project. Thank you for your aid in completing this work, also in times that situations were more complex and demanded extra time and effort.

Mama en papa, bedankt voor jullie blijvende ondersteuning, ook tijdens de moeilijkere periodes. Bedankt voor mij de luxe van extra tijd te geven die ik nodig had om dit werk tot een goed einde te brengen.

Lars, mijn roommate, bedankt voor alle leuke tijden. Green Day nummers spelen op gitaar tijdens de pauze was slechts een van de momenten die ik nooit zal vergeten. Dank je voor je aanwezigheid, hulp en het vele advies dat ik van jou gekregen heb.

Hanna, thanks for being there. Thank you for your advice on different levels of work and also on life, and for being my “favourite post-doc”.

Federico, thanks for you friendship next to being my main method consultant for the first chapter. Thank you for the countless times I was able to get your advice with methodology that was above my level of understanding. Giancarlo, thank you for your assistance with the methods of the last chapter. You are a kind and understanding person and without your work and assistance this thesis would not have been completed.

Gonny en Dries, bedankt voor alle leuke tijden die we hebben meegemaakt. Bedankt voor jullie vriendschap en ondersteuning. Ik wil jullie naast Petra, Anniek en Vera ook bedanken voor de mooie reis naar Californië. Samen met bovengenoemde wil ik ook Teresa en Jeanette bedanken voor de ontelbare leuke lunch-momenten en coffee breaks, waarbij geen enkel gespreksonderwerp werd gespaard.

Milene, Michelle, Nick, Roberta, Anke and all the others of the Auditory Research Group. I would like to thank you for your scientific input during the many meetings. Rainer, thank you for your assistance, as well as for your highly contaminating passion for science. Annemie en Christl, bedankt voor de hulp bij alle stappen van het afwerken van het proefschrift. Marin en Aline, bedankt voor alle leuke feestjes. The following people I would like to thank just for being there: Bernadette, Henk, Fabrizio, Vincent, Bert, Francesco, Alex, Christianne, Nina, Tom, Job, Riny, Pim, Sven, Martin, Michael, Alard, Riny, Bettina, Alessia, Amanda, Peter, Joel, Judith, Armin and the people I forgot.

Erik en Elsbeth, bedankt voor alle geweldige momenten. Ook nog dank aan Johan, Jonas, Don, Jan, Florian, Bjorn, Philipe, Kristof, Christine, Rens, Jochen en Dominique.

Curriculum vitae

Noël Staeren was born on January 6th 1980 in Bilzen, Belgium. In 1998 he graduated from high school at the Technisch instituut Sint-Josef in Bilzen, Belgium and enrolled at the Faculty of Psychology at Maastricht University, where he also worked as a teaching assistant. After obtaining the propedeutical degree in psychology, he enrolled in the neuropsychology master track. In 2001 he attended optional courses in computer science at the department of knowledge engineering of the same university, where he studied for three years. As part of the knowledge engineering track, he followed a summer course at the computer science department of Baylor University, Texas. After obtaining the propedeutical degree and completing selected master courses in knowledge engineering, he started his research practical at the Cognitive Neuroscience department of the faculty of psychology under the supervision of dr. Elia Formisano, titled "Comparing methods for analysing effective connectivity in fMRI: Application to a spatial imagery experiment". In 2005 he obtained his master's degree in Psychology at Maastricht University with a specialization in Cognitive Neuroscience. He worked on his PhD project at the Cognitive Neuroscience department at Maastricht University from 2005-2014.

Publications

Renvall, H., Staeren, N., Siep, N., Esposito, F., Jensen, O., Formisano, E. (2012) Of cats and women: Temporal dynamics in the right temporoparietal cortex reflect auditory categorical processing of vocalizations. Neuroimage 62(3), 1877-1883.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., Formisano, E. (2009) Sound categories are represented as distributed patterns in the human auditory cortex. Current Biology 19(6), 498-502.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E. (2008) Combining multivariate voxel selection and Support Vector Machines for mapping and classification of fMRI spatial patterns. Neuroimage 43(1), 44-58.

Sack, A., Jacobs, C., De Martino, F., Staeren, N., Goebel, R., Formisano, E. (2008) Dynamic Premotor-To-Parietal Interactions during Spatial Imagery. The Journal of Neuroscience 28(34), 8417– 8429.

Smolders, A., De Martino, F., Staeren, N., Scheunders, P., Sijbers, J., Goebel, R., Formisano, E. (2007) Dissecting cognitive stages with time-resolved fMRI data: a comparison of fuzzy

clustering and independent component analysis. Magnetic Resonance Imaging 25(6), 860 – 868.

Poster presentations

Human Brain Mapping 2008, Melbourne, Australia - Renvall, H., Staeren, N., Siep, N., Jensen, O., Formisano, E. Neural correlates of auditory categorical perception revealed by magnetoencephalography.

Auditory Cortex 2006, Nothingham, United Kingdom - Staeren, N., Renvall, H., De Martino, F., Goebel, R., Formisano, E. Distributed representations of sound categories in the human temporal lobe.

Human Brain mapping 2006, Florence, Italy - Staeren, N., Renvall, H., De Martino, F., Goebel, R., Formisano, E. Distributed representations of sound categories in the human temporal lobe.

Human Brain Mapping 2005, Toronto, Canada. - Staeren, N., Roebroeck, A., Sack, A., Goebel, R., Formisano, E. Mapping directed cortical interactions during visuospatial imagery using fMRI mental chronometry and Granger Causality.

In preparation

Staeren, N., Renvall, H., Schreiner, C., Walter, A., Goebel, R., Formisano, E. (in preparation). Cortical processing of spatial cues in natural auditory scenes.

Staeren, N*, Valente, G.*, Formisano, E*. (in preparation, *equal contribution). Brain-based un-mixing of vocal and instrumental streams during music listening