Processing of natural sounds and scenes in the human brain

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.