Activity Associated with Stream Segregation in Human Auditory Cortex is Similar for Spatial and Pitch Cues

Stefan Schadwinkel1,2 and Alexander Gutschalk1,2

1Department of Neurology, Ruprecht-Karls-Universität Heidelberg, 69120 Heidelberg, Germany and 2Department of Neuroradiology, Ruprecht-Karls-Universität Heidelberg, 69120 Heidelberg, Germany

Address correspondence to Stefan Schadwinkel, Department of Neurology, University of Heidelberg, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany. Email: stefan.schadwinkel@med.uni-heidelberg.de.

Stream segregation is a perceptual mechanism by which the brain segregates information from multiple sound sources in our environment and assigns them to distinct auditory streams. Examples for streaming cues are differences in frequency spectrum, pitch, or space, and potential neural correlates for streaming based on spectral and pitch cues have been identified in the auditory cortex. Here, magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) were used to evaluate if response enhancement in auditory cortex associated with streaming represents a general pattern that is independent of the stimulus cue. Interaural time differences (ITDs) were used as a spatial streaming cue and were compared with streaming based on fundamental frequency ($f_0$) differences. The MEG results showed enhancement of the $P_{2m}$ after 60–90 ms that was similar during streaming based on ITD and pitch. Sustained fMRI activity was enhanced at identical sites in Heschl’s gyrus and planum temporale for both cues; no topographical specificity for space or pitch was found for the streaming-associated enhancement. These results support the hypothesis of an early convergence of the neural representation for auditory streams that is independent of the acoustic cue that the streaming is based on.

Keywords: auditory scene analysis, functional magnetic resonance imaging, ITD, magnetoencephalography, stream segregation

Introduction

In order to communicate in busy auditory environments, our brain needs to separate concurrent and overlapping information from multiple sound sources into distinct perceptual entities. Auditory stream segregation (or streaming) is considered to be a fundamental mechanism to solve this problem (Bregman 1990; Moore and Gockel 2002; Carlyon 2004), and a number of microelectrode studies in animal models (Bee and Klump 2004; Fishman et al. 2004; Micheyl et al. 2005; Pressnitzer et al. 2008; Elhilali et al. 2009; Itatani and Klump 2009) and neuroimaging studies in human listeners (Sussman et al. 1999; Deike et al. 2004; Cusack 2005; Gutschalk et al. 2005; Snyder et al. 2006; Gutschalk, Oxenham, et al. 2007; Wilson et al. 2007) have lately tackled its neural underpinnings.

The animal studies used pure tones and suggested that 1) the separation of streams into distinct neural populations along the tonotopic axis and 2) the associated frequency-selective adaptation are 2 elementary mechanisms for perceptual streaming (Bee and Klump 2004; Fishman et al. 2004; Micheyl et al. 2005; Pressnitzer et al. 2008). Recent data suggest that the concept of a separation into different neural populations can be transferred to amplitude modulation as a streaming cue (Itatani and Klump 2009).

Human imaging studies (Gutschalk et al. 2005; Wilson et al. 2007) also tested the relationship between streaming and (frequency) selective adaptation in the auditory cortex. Specifically, the amount of adaptation in magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) closely covaried with the perceptual organization, such that there was more interaction between tones in one stream compared with tones organized into separate streams. Thereby, the amount of adaptation reflected the dominant time interval perception and its covariation with the perceptual organization. Analogous covariation of streaming perception and selective adaptation was demonstrated for nonspectral pitch (Gutschalk, Oxenham, et al. 2007) and produced similar results as the previous setups using spectral streaming cues in the auditory core and belt areas. These results suggest that streaming based on various cues (Moore and Gockel 2002) may rely on similar neural mechanisms in the auditory cortex.

It is currently unclear whether stimulus specificity in the auditory belt cortex is as modular as in the visual system, but a number of topographic specializations within the auditory cortex have lately been reported: One is specificity for pitch lateral from the primary auditory cortex (Griffiths et al. 1998; Gutschalk et al. 2002; Patterson et al. 2002; Warren and Griffiths 2003; Penagos et al. 2004; Bendor and Wang 2005). Another one is the separation of an anterior “what” pathway from a more posterior “where” pathway (Alain et al. 1998; Rauschecker and Tian 2000; Tian et al. 2001; Zatorre et al. 2002; Arnott et al. 2004; Ahveninen et al. 2006; Barrett and Hall 2006; Lomber and Malhotra 2008), in the anterior and posterior belt fields of the auditory cortex.

One possibility was therefore that the sites where a release from selective adaptation associated with streaming is found vary with the stimulus cue that the streaming is based on. Alternatively, streaming-associated adaptation in the auditory cortex might reflect a more general representation of auditory perceptual organization that is independent of the individual streaming cue, in which case no topographic variation with the stimulus cue would be expected.

In this study, we evaluated the generality versus specificity of MEG and fMRI activity enhancement during streaming across stimulus cues. For this purpose, we studied whether streaming-associated selective adaptation in auditory cortex, identified in previous studies (Gutschalk et al. 2005; Gutschalk, Oxenham, et al. 2007; Wilson et al. 2007), showed topographic specificity depending on whether the streaming was based on pitch or space. The stimuli used elicited a high probability of streaming either based on a difference in the $f_0$, producing streams with different pitches, or based on a difference in the interaural time.
difference (ITD; Hartmann and Johnson 1991; Bochinke and Phillips 2005), producing streams with different lateralization.

Our hypothesis was, first, that the same modulation of neural activity in the auditory cortex observed for pitch and spectral cues would also be found for the spatial cue. Alternatively, if release from selective adaptation associated with streaming was related to topographic stimulus specificity, we expected that the pitch cue would rather modulate activity in (lateral) Heschl’s gyrus, whereas the spatial cue should rather modulate activity in the planum temporale (Patterson et al. 2002; Warren and Griffiths 2003; Deouell et al. 2007).

### Materials and Methods

#### Listeners

Twelve listeners (8 female) aged between 22 and 41 years (mean 28) participated in both MEG and fMRI measurements. None of the listeners had a history of central or peripheral hearing disorder. Three additional listeners took part in pilot psychophysical measurements in a quiet booth, one of them participated in a pilot MEG measurement, the results of which were used to select suitable stimulus parameters for fMRI and MEG. All subjects gave written informed consent to participate in the study after the study was explained to them. The study was performed in accordance with the Declaration of Helsinki, and the Ethics Committee of the Medical Faculty of Ruprecht-Karls-University Heidelberg approved the study.

#### Stimulation

The stimulus sequences comprised harmonic tone complexes, each 125 ms in duration (including 10-ms raised-cosine onset and offset ramps); all partials had the same amplitude and started in sine phase. The tone complexes were generated digitally at a sampling rate of 48 kHz (used during MEG measurements) and 32 kHz (used during fMRI measurements) with 16-bit resolution and were digitally low-pass filtered (sixth-order Butterworth filter with zero phase shift) at 5 kHz. This cutoff frequency was selected because the transfer functions of the sound delivery systems used in the MEG and fMRI measurements were reasonably linear below. No lower cutoff frequency to avoid that the lower harmonics can be perceptually resolved was used because ITD-based discrimination deteriorates above 1500 Hz (Moore 2004) and a high-pass filter of about 2 kHz would have been required to avoid the lower harmonics from being perceptually resolved. The pitch of the tones can therefore be based on both spectral and temporal cues, which is different from the experiment by Gutschalk, Oxenham, et al. (2007), where spectral cues were avoided.

The complex tones were temporally arranged into a repeating ABBB pattern, where A and B represent tones that differed either in f0 or in ITD, according to their respective condition, with no silent gaps between consecutive tones. This pattern was chosen because it can be similarly used in both MEG and fMRI (Gutschalk et al. 2007).

In our MEG measurements, we planned to focus on responses to the A tones because the percept-dependent differences in perceived interstimulus interval (ISI) were larger for A than B tones. Therefore, the A tones were held constant at an f0 of 180 Hz and a lateralization to the left by an ITD of about ~700 μs. As we used different sampling rates for the stimuli during MEG and fMRI measurements due to technical reasons, a rounding error displaced the 2 conditions by one sample, which caused the resulting ITD to be ~708.3 μs at 48 kHz (used during MEG) and ~687.5 μs at 32 kHz (used during fMRI). Unfortunately, the measurements had already been completed by the time the error was discovered. We think that this error is negligible, though, because the just noticeable difference between ITDs at this maximum physiological lateralization is above 60 μs (for all frequencies below 5000 Hz) (Moore 2004), while the difference between the 2 sample rates was only 20.8 μs.

The parameters of the B tones were varied according to 1 of 3 conditions: 1) a nonstreaming (control) condition with identical parameters for the B tones and the A tones (NO condition), 2) an ITD-based streaming condition where the B tones had an f0 identical to the A tones but were lateralized to the right, using an ITD of 708.3 μs at 48 kHz and 687.5 μs at 32 kHz for MEG and fMRI, respectively (ITD condition), and 3) an f0-based streaming condition where the B tones had an f0 of 120.1 Hz (a difference of 7 semitones) and the same ITD as the A tones (F0 condition).

This frequency difference was chosen based on pilot studies that showed that the amount of streaming was similar for this condition compared with the ITD-based condition with lateralization of A to the left and B to the right, each by 700 μs. The stimuli were 32-s sequences of one of these constant conditions. Condition was varied pseudorandomly across sequences.

#### Procedures

Before the measurements, listeners were familiarized with the stimuli and task. The objective was explained by the first author, along with a schematic drawing explaining what was meant by “1 stream” and “2 streams”, and they performed a few practice trials. During the actual measurements, the sequences corresponding to the 3 conditions were presented 20 times each in random order. While listening to the sequences during the measurements, listeners assumed a one-stream percept at the beginning of each 32-s sequence and indicated as soon as possible when the percept switched into a 2-stream percept and then again after each change in percept (from 1 to 2 streams and vice versa) by pressing 1 of 2 keys, corresponding to the 2 percepts.

During fMRI, the sequences were presented in 5 runs, each comprising 4 repetitions of each condition, i.e., twelve 32-s sequences per run. A blocked stimulus design was used in which the presentation of each 32-s sequence was separated by 34 s of silence. The time between runs was usually 30–60 s. Sound presentation was diotic via electrodynamic headphones (MR Confon; MR confon GmbH). Listeners were instructed to focus on the tones and indicate their percept using a hand response box (LUMItouch fMRI optical response keypad; Photon Control). Responses were recorded using Presentation (Neurobehavioral Systems).

During MEG, sounds were presented using ER-2 tube phones (Etymotic Research), shielded by custom-made 2-layered copper boxes, with 90-cm-long plastic tubing. The 32-s sequences were separated by 8 s of silence. Each condition was presented 20 times in random order yielding a total of 1280 repetitions of the A tones per condition (each 32-s sequence contained 64 ABBB quadruplets). Listeners reported their percept continuously as described above using a modified MEG-compatible standard computer mouse.

During MEG and fMRI, the stimuli were presented at an overall level of 70 dB sound pressure level (SPL). All tones were normalized to the same root-mean-square value. The SPL at the ear was measured using a sound-level meter (Brüel&Kjær) for the MEG sound delivery equipment. The sound-presentation system used during the fMRI sessions allowed for digital input of the stimuli and provides a verified output level, as determined by the manufacturer at the installation of the system in the magnetic resonance imaging (MRI) scanner, which was used to adjust the overall SPL to 70 dB SPL.

#### Data Acquisition

MEG data were acquired using a 3T scanner (Magnemot Trio; Siemens) equipped with an 8-channel phased-array head coil. Functional imaging was performed using an echo-planar imaging sequence (gradient echo; time echo 30 ms; flip angle 90°; in-plane resolution 64 × 64; field of view [FOV] 200 × 200 mm; 12 slices; slice thickness 3.1 mm; gap 1.023 mm). The volume for functional imaging was chosen as 12 coronal slices (perpendicular to the Sylvian fissure, covering the auditory cortex from the posterior end of planum temporale to the anterior aspect of the superior temporal gyrus, including the complete Heschl’s gyrus [first and second when there were 2]) in both hemispheres. Images of the 12-slice volume were acquired in brief 1-s clusters separated by a 7-s quiet interval to decrease the interference of the imager noise with the auditory stimulation (Edmister et al. 1999; Hall et al. 1999). To allow reconstruction of the blood oxygen level-dependent (BOLD) signal time course with 2-s resolution, the stimulus presentation was repetitively delayed relative to image acquisition by 2 s using 34 s of silence between sequences.
The sequences per condition were pseudorandomized in respect to the silence intervals to acquire each of the 4 repetitions per condition within each run with a delay of 0, 2, 4, and 6 s, respectively. For coregistration, a structural magnetization prepared rapid gradient echo (MP-RAGE) image of the whole head (sagittal in-plane resolution 256 × 256; FOV 256 × 256 mm; 128 slices; slice thickness 1.3 mm) and a high-resolution T2-weighted structural image (in-plane resolution 512 × 512; FOV 200 × 200 mm; 12 slices; slice thickness 3.1 mm) of the same volume as the functional images were acquired. For 11 of 12 listeners, 2 additional MP-RAGE images of the whole head (sagittal in-plane resolution 256 × 256; FOV 256 × 256 mm; 128 slices; slice thickness 1.3 mm) were acquired using a standard bird-cage head coil, in order to gain a higher quality of the cortical reconstruction.

MEG data were acquired with a Neuromag-122 whole-head MEG system (Elekta Neuromag Oy) (Ahonen et al. 1993) in a magnetically shielded room (IMEDECO). This system comprises 61 dual-channel planar first-order gradiometers in a helmet-shaped configuration, measuring the magnetic field gradient in 2 orthogonal tangential directions. The data were recorded at a sampling rate of 1000 Hz and filtered online with a bandwidth of 0.01–350 Hz. Before imaging, 4 position indicator coils were fixed to the subject’s head and the position of the coils and 32 additional points on the head surface were digitized relative to landmarks on the head surface. The positions of the head position coils relative to the MEG device were determined at the beginning of the recording session.

Data Analysis

The fMRI data were motion corrected using tools from the AFNI (National Institute of Mental Health) software package (Cox and Jesmanowicz 1999). No further preprocessing was applied to the data prior to detecting activation.

Activity was detected by a general linear model using a single-gamma ($\alpha = 2.25$, $\tau = 1.25$; Dale and Buckner 1997) hemodynamic response function convolved with a boxcar function using the Functional Analysis Stream (FSFAST) from the FreeSurfer software package (Athinoula A. Martinos Center for Biomedical Imaging). Drift components were modeled with second-order polynomial regressors. Motion correction parameters were included as regressors in the model. The remaining noise was modeled as time-invariant linear AR(1) process. Activation maps were derived by a random-effects multi-subject analysis (not corrected for multiple comparisons). The multi-subject analysis was performed within the cortical surface space (Fischl et al. 1999), coregistered to the FreeSurfer brain template (fsaverage). For this purpose, the whole-head images were processed with FreeSurfer to create an inflated projection of the cortical surface (Dale et al. 1999; Fischl et al. 1999, 2001; Segonne et al. 2004). Patches of the superior temporal plane were computationally “snipped” from the inflated template surface that is distributed with FreeSurfer and were flattened for display. The analysis was restricted to the auditory cortex. Coregistration between functional and structural data, and between structural data from different imaging runs, was done using BET (Smith 2002) and FLIRT (Jenkinson and Smith 2001) from the FSL (FMRIB) software package (Smith et al. 2004).

Activation time courses were derived as follows: The respective group activation maps were projected back to the surface space of each subject. The reconstructed, intensity normalized, and averaged (across sequence repetitions per condition) BOLD response time courses in the volume space were projected into the surface space of the corresponding subject. From these projections, the time course of each ‘active’ vertex was determined and averaged per subject. These per subject time courses were then low-pass filtered (second-order Butterworth filter with zero phase shift) at 0.125 Hz and averaged to obtain an average group activation time course. For each subject and condition, a set of physiological basis functions (Harms and Melcher 2003) was fit to the average waveform, using nonlinear least-squares regression. The obtained regression coefficients were then used to compute the waveshape index (Harms and Melcher 2003) for each waveform. Amplitudes of the onset peak (maximum from 2 to 12 s), early sustained part (average from 12 to 22 s), late sustained part (average from 22 to 32 s), and the offset peak (maximum from 34 to 40 s) were measured in the individual time courses for both hemispheres and all 3 conditions. The amplitude and waveshape index values were separately submitted to a repeated measures analysis of variance (ANOVA), calculated using a multivariate linear modeling procedure in R (The R Foundation for Statistical Computing) with the independent variables ‘condition’, ‘hemisphere’, and ‘component’, the latter only for the amplitude values. Amplitudes from 2 different contrasts were submitted to this procedure with the additional independent variable “contrast.” Significance levels were not corrected for multiple comparisons. The Greenhouse-Geisser sphericity correction was applied to the degrees of freedom for univariate tests where appropriate. All reported $p$ values were based on the reduced degrees of freedom, although the original degrees of freedom are reported.

The MEG data were analyzed with BESA (MEGIS Software). The 1280 ABBB quadruplet repetitions that occurred during each condition were submitted to the BESA artifact scan tool to automatically remove artifact-contaminated epochs (average rejection rate 3.5%, range 1.6–10.2%). The remaining epochs (excluding the first epoch of each 32-s sequence) were then averaged; time intervals began 125 ms before and ended 500 ms after the onset of each quadruplet. A baseline, calculated in an interval of 125 ms before the onset of the A tone, was subtracted from the response.

A source analysis (Scherg and von Cramon 1986; Scherg 1990) was performed separately for each subject. It used a spherical head model and assumed 2 dipole sources (one in each auditory cortex). The dipole space was individually coregistered to the T1-weighted whole-head image for each subject using BrainVoyager (Brain Innovation B.V.) to obtain Talairach coordinates of dipole positions later on. For the source analysis, the data were band-pass filtered to 0.01–30 Hz (first- and second-order Butterworth filter with zero phase shift, respectively) to avoid high-frequency artifacts. The data of one subject exhibited a slow negativity causing displacement of the $P_m$; these data were band-pass filtered to 5–30 Hz (first- and second-order Butterworth filter with zero phase shift, respectively) for the purpose of dipole fitting. As a control, the data of this subject were excluded from the group averages. This did not lead to relevant differences in the results; therefore, the data from this subject were included in the results reported below. Dipoles were fit to the rising flank of the $P_m$ elicited by the A tone for each streaming condition alone. The Talairach coordinates of both dipoles were then submitted to the same ANOVA procedure as used for the fMRI data (see above) with the independent variables condition, hemisphere, and “dimension”. For the purposes of obtaining source waveforms, the dipoles were fit to the rising flank of the $P_m$ elicited by the A tone in a superposition of the 2 streaming conditions.

Principal component analysis (PCA) was used to correct for low-frequency drifts. The data were low-pass filtered at 3 Hz (second-order Butterworth filter with zero phase shift), and a PCA component was obtained. After the last component was included in the model, the model was then used as a spatial filter to derive unfiltered source waveforms across all 3 conditions. These source waveforms were then low-pass filtered (fourth-order Butterworth filter with zero phase shift) at 30 Hz and averaged across subjects with MATLAB (Mathworks).

The dynamics of the MEG data during the 32-s stimulus block were additionally reconstructed as follows: Each of the 64 ABBB quadruplet positions comprised in a 32-s sequence was individually averaged across the 20 repetitions per condition. Another average with a higher signal-to-noise ratio (SNR) was computed by pooling each subsequent 4 repetitions of the ABBB quadruplet prior to averaging, which allowed to average 80 epochs per time position. The dipole models (including the PCA component) from the main analysis were used as a spatial filter to derive unfiltered source waveforms for each time position, condition, and hemisphere from both averages. Using MATLAB, these waveforms were then low-pass filtered at 30 Hz (zero phase shift, fourth-order Butterworth filter). As these waveforms exhibited more noise and drifts due to the small number of averaged epochs, peak-to-peak amplitudes were used to assess the amplitudes of the $P_m$ and $P_{2m}$ peaks. Three amplitude values were evaluated: 1) between the negative peak before the $P_m$ (referred to here as $N_{m}$) and the $P_m$, 2) between the $P_m$ and the $N_{m}$, and 3) between the $N_{m}$ and the $P_{2m}$. Three waveforms exhibited a strong drift: For each of these
3 waveforms, another PCA component was individually added as explained above.

Peak amplitudes and latencies were determined by finding turning points using the first derivative. Turning points with maximum (for \( P_1 \) and \( P_2 \)) or minimum (for \( N_0 \) and \( N_1 \)) amplitude in fixed time intervals were taken as the respective peak amplitudes. Additional consistency constraints were applied to the fixed time intervals. The \( P_2 \) peak was determined first, then the \( P_1 \) peak was required to be at maximum the penultimate turning point before the \( P_2 \) peak, and finally the \( N_1 \) peak was determined and restricted to turning points between the \( P_2 \) and \( P_1 \) peaks. The intervals for this procedure (respective to A-tone onset) were ~50 to 75 ms for the \( N_0 \) peak, 20–150 ms for the \( P_1 \) peak, 45–240 ms for the \( N_1 \) peak, and 135–290 ms for the \( P_2 \) peak. Peaks and latencies of the overall source waveforms were submitted to statistical analysis with the independent variables condition and hemisphere using the same ANOVA procedure as used with the fMRI data (see above).

Behavioral responses were used to determine the probability of perceiving segregated streams after every ABBB quadruplet (500 ms) over the course of the 32-s stimulus block. These time courses were averaged across repeated presentations, separate for each measurement (MEG and fMRI). The probabilities to perceive streaming for each 4 subsequent quadruplets were pooled to yield a resolution of 2 s. For statistical analysis, the data (increased by one for computational reasons) were transformed using a Box-Cox power transform to stabilize variances. The \( \lambda \) parameter of the transform was estimated based on a log-likelihood profile computed with R for the respective linear model. These data were then submitted to the ANOVA procedure as described above for the fMRI data, with the independent variables condition, ‘measurement’, and ‘time’.

**Results**

**Behavioral Data**

Figure 1 shows the average percentage of 2-stream responses as a function of time for all conditions across subjects. For both the ITD and the F0 conditions, streaming built up reliably in the first 5–8 s and remained at a constantly high level thereafter. In contrast, no significant streaming perception was observed for the control (NO) condition. The difference between the streaming conditions and the control condition was significant in an overall ANOVA, including both measurements (MEG and fMRI), all 3 conditions, and the full-time segment \((F_{2,11} = 406.95, P < 10^{-11})\). The buildup is reflected by a significant time effect \((F_{1,5,105} = 10.915, P < 0.05)\) and condition \(\times\) time interaction \((F_{30,210} = 16.634, P < 0.001)\). There was somewhat less streaming in fMRI compared with MEG \((F_{1,7} = 5.7335, P < 0.05)\), while the condition \(\times\) measurement interaction was not significant in this analysis \((F_{2,11} = 1.5644, P > 0.05)\).

Further, an analysis including only the 2 streaming conditions failed to detect a significant difference between the ITD and F0 conditions \((F_{1,7} = 3.2319, P > 0.05)\). One subject answered during MEG using a different response strategy (ignoring the 8-s silent pauses) that disrupted the full reconstruction of the time points where perceptual changes due to streaming occurred. After discussing the behavioral responses with the subject, performance during fMRI revealed reliable detection of streaming percepts, and thus, the MEG data were included in the analysis. The responses of 3 subjects during fMRI were lost due to a technical malfunction. The behavioral data from these subjects during MEG, however, revealed reliable responses, and they were also included in the further analysis. All available data \((n_{\text{MEG}} = 11, n_{\text{fMRI}} = 9)\) were used for the plot in Figure 1, while only the data available from both measurements \((n = 8)\) were used for the statistical analysis above. Plotting the responses from only these 8 subjects did not produce relevant differences.

**Magnetoecephalography**

A \( P_1 \) was consistently elicited by the A tones of ABBB sequences for all conditions and had a mean latency between 66 and 90 ms (across conditions).

The average dipole position fitted to the \( P_1 \) was located in the anterolateral part of Heschl’s gyrus in both hemispheres and is indicated by white circles in Figure 5 (Talairach coordinates \([x; y; z \pm \text{standard errors}]; \text{left}: [-46.1 \pm 2.5; -17.1 \pm 2.7; 1.4 \pm 2.3]; \text{right}: [42.7 \pm 3.0; -15.4 \pm 2.5; 0.8 \pm 1.8]). Dipoles fitted to the \( P_1 \) for each streaming condition individually did not differ significantly in position (ITD vs. F0: \( F_{1,11} = 0.8859, P > 0.05)\). The source waveforms corresponding to the dipoles are shown in Figure 2. A schema of the stimulus (ABBB tone pattern) is shown at the bottom. The mean amplitudes of the major deflections in the source waveforms (i.e., \( P_1 \), \( N_1 \), and \( P_2 \)) are depicted in Figure 3 for each condition.

The amplitude of the \( P_1 \) changed significantly with condition \((F_{2,22} = 13.848, \ P < 0.01)\). Consistent with our hypothesis, the amplitude increased for both streaming conditions compared with the nonstreaming condition (ITD: \( F_{1,11} = 21.799, P < 0.001; \text{F0}: F_{1,11} = 12.684, P < 0.01)\). There was no difference between the two streaming conditions \((F_{1,11} = 0.173, P > 0.05)\). There were no significant hemisphere effects in the ANOVA.

Statistical analysis of the \( P_1 \) latencies revealed a significant main effect of condition \((F_{2,22} = 8.9791, P < 0.01)\). This effect was mainly due to longer \( P_1 \) latencies (lat\(t \) and lat\(g \) for left and right hemisphere, respectively) in the streaming compared with the NO condition (ITD: lat\(t \) = 86.7 ± 5.3 ms, lat\(g \) = 72.4 ± 4.3 ms; F0: lat\(t \) = 89.5 ± 4.1 ms, lat\(g \) = 84 ± 4.6 ms; NO: lat\(t \) = 69.1 ± 6.3 ms, lat\(g \) = 66.5 ± 4 ms). Average response latencies were slightly earlier in right auditory cortex, but the difference was only significant for the ITD condition (ITD: L vs. R, \( t = 2.4707, P < 0.05\)) and did not produce significant main effects or interactions in the ANOVA.

Following the \( P_1 \), an \( N_1 \) component was consistently observed for the NO and ITD conditions. A trough during the same time interval was also found for the F0 condition but was elevated by an overlapping slow positivity. For this reason, the \( N_1 \) was not analyzed in detail. Mean latencies (across conditions) of the \( N_1 \) were between 125 and 144 ms.
A second positive deflection ($P_2m$) occurred at a mean latency (across conditions) of 190–204 ms. The locations of dipoles fit to the $P_2m$ were not significantly different from the $P_1m$ dipoles ($F_{1,11} = 0.333, P > 0.05$); the $P_1m$ dipoles were therefore used for the analysis reported below. For both streaming conditions, the $P_2m$ was increased in amplitude in comparison to the NO condition (ITD: $F_{1,11} = 26.54, P < 0.001$; F0: $F_{1,11} = 27.792, P < 0.001$). The $P_2m$ was significantly more pronounced in the F0 condition ($F_{1,11} = 17.991, P < 0.001$) and was somewhat larger in the right hemisphere (hemisphere: $F_{1,11} = 5.057, P < 0.05$; condition × hemisphere: $F_{1,11} = 1.6814, P > 0.05$).

Note, however, that the $P_2m$ component evoked by the ABBB sequences is not a “pure” $P_2m$ but a blend of a $P_2m$ evoked by the A tones and a $P_1m$ component elicited by the first B tone. This assumption was confirmed by an additional experiment in a single subject, using control stimuli consisting only of A-- and -BBB sequences (each “—” indicating a 125-ms silence period).

To evaluate the evolution of the different evoked response components over time, the data were averaged in steps of 1 or 4 quadruplets and transformed into source waveforms as above (see Materials and Methods for details). To partly compensate for the higher low-frequency noise level in these data, only peak-to-peak amplitudes were analyzed, including $N_0m$-$P_1m$ (Fig. 4, top row), $P_1m$-$N_1m$ (Fig. 4, middle row), and $N_1m$-$P_2m$ (Fig. 4, bottom row). The analysis revealed a strong adaptation within the first 1 to 4 quadruplets for the $P_1m$-$N_1m$ and the $N_1m$-$P_2m$ amplitude, which might possibly be related to the drop in signal intensity after the initial BOLD transient (see below). No amplitude reduction was observed for the $N_0m$-$P_1m$. This lack of $P_1m$ adaptation might either be due to
cancellation between the $P_1m$ and $N_1m$ waves (the $P_1m$ is often small when the $N_1m$ is prominent because the $N_1m$ onset overlaps the $P_1m$ peak) or reflect a buildup of the $P_1m$, as suggested by Snyder et al. (2006). In the subsequent interval, the $N_0m$-$P_1m$ and $N_1m$-$P_2m$ for the F0 and ITD conditions were generally higher as for the nonstreaming control.

**Functional magnetic resonance imaging**

Figure 5 shows areas of activation in the auditory cortex on an inflated and flattened cortex patch taken from the averaged brain template of FreeSurfer. The ITD + F0 versus NO contrast reveals that multiple regions of the auditory cortex were activated by both streaming conditions, including Heschl’s gyrus, the planum temporale, and parts of the superior temporal gyrus. The individual streaming contrasts (ITD vs. NO and F0 vs. NO) reveal that both conditions show activity in qualitatively the same anatomical regions in anterolateral Heschl’s gyrus and planum temporale. Although activity was somewhat more extended for the $f_0$-based streaming, the center of activity overlapped for both conditions and activity was even more extended for the combined ITD + F0 versus NO contrast. There was only a small and sketchy area where activation was significantly stronger for the F0 compared with the ITD condition (F0 vs. ITD). Conversely, there was no area within auditory cortex where activity was significantly higher for the ITD condition (ITD vs. F0; data not shown). Lowering the statistical threshold for the F0 versus ITD contrast to $P < 0.01$ (the white area in the outmost patches in Fig. 5b) increased the area but did not reveal activity at sites distinct from those active in the ITD + F0 versus NO contrast.

Figure 6 shows the time course of the BOLD response, averaged across subjects and active vertices for the ITD + F0
versus NO contrast ($P < 0.001$, see Fig. 5) and separated for hemispheres. In all 3 conditions, the BOLD response starts with a transient response that peaks at about 6 s after sequence onset, followed by a trough or sustained part and another peak after sequence offset.

Our hypothesis was that the BOLD activity was more sustained during streaming, due to a decrease in perceived tone repetition rate (Harms and Melcher 2002; Gutschalk, Oxenham, et al. 2007; Wilson et al. 2007). Enhanced sustained BOLD activity has already been suggested by the difference contrast reported above, for which a sustained regressor was used. The effect was evaluated in more detail with 2 other measures: first, the waveshape index suggested by Harms and Melcher (2003), which was computed for the response time courses of all 3 conditions (Fig. 7). The waveshape index ranges from 0 to 1; it is small when the signal is sustained for the whole duration of a block, and it is large for a phasic response consisting of transients at the beginning and ending. For the present data, the waveshape index was smaller for the streaming conditions compared with the nonstreaming condition. In the ANOVA, there was a significant main effect of condition ($F_{2,22} = 12.087, P < 0.01$) and there was no significant effect or interaction for hemisphere. There was a significant difference between each of the streaming conditions and the NO condition (NO vs. ITD: $F_{1,11} = 7.4393, P < 0.05$; NO vs. F0: $F_{1,11} = 16.95, P < 0.01$) but also a small difference between the 2 streaming conditions ($F_{1,11} = 8.9311, P < 0.05$). Thus, the waveshape index reveals a significant change from phasic (during the NO condition) to more sustained response waveshape during streaming (Wilson et al. 2007), which was more pronounced for the F0 condition. Note that the time courses of the BOLD response in the area detected by the F0 versus ITD contrast were highly similar to the time courses obtained from the area detected by the ITD + F0 versus NO contrast for all 3 conditions and produced the same pattern of significant effects in waveshape index and
Thus, the area detected by the F0 versus ITD contrast reflects the general finding of higher BOLD amplitude for the F0 compared with the ITD condition but does not reveal a site specifically activated by streaming based on f0-differences.

Second, amplitudes of the onset transient, early (12–22 s), and late (22–32 s) sustained parts and the offset transient were measured. For both streaming conditions, the early sustained amplitude increased significantly (NO vs. ITD: F1,11 = 11.544, P < 0.01; NO vs. F0: F1,11 = 52.39, P < 10^{-3}). The overall sustained level was slightly higher during the F0 than the ITD condition, during both early and late time intervals (early: F1,11 = 13.029, P < 0.01; late: F1,11 = 4.8964, P < 0.05). For the late interval, sustained levels were not significantly different between the nonstreaming and streaming conditions (NO vs. ITD: F1,11 = 1.0621, P > 0.05; NO vs. F0: F1,11 = 4.4207, P > 0.05). The onset transient increased significantly with streaming (NO vs. ITD: F1,11 = 11.758, P < 0.01; NO vs. F0: F1,11 = 30.552, P < 0.001) and was not different between streaming conditions (ITD vs. F0: F1,11 = 4.1589, P > 0.05). No significant effects were found for the OFF peak following the stimulus block.

Figure 5. (a) Significant auditory cortex activation mapped on the flattened cortical surface of the superior temporal plane (random-effects group analysis; n = 12; cutoff P < 0.001, not corrected for multiple comparisons). The location of Heschl’s gyrus (HG), the Planum temporale (PT), and the superior temporal gyrus (STG) are indicated on the top left patch. The average position of the P1m dipoles in each auditory cortex is indicated by a white-filled black circle. The first row shows the contrast of each condition with the silent baseline. The nonstreaming control condition (NO) shows only weak activation, while both streaming conditions (F0 and ITD) produce extended sustained BOLD activation in and surrounding the complete Heschl’s gyrus. The second row shows contrasts of both streaming conditions versus the nonstreaming control. The average of both conditions, which was used for the ROI analysis, is shown on the left. (b) Overlay of the activity of the ITD versus NO (blue) and F0 versus NO (red) contrasts and their overlap (green). The activation from the F0 versus ITD contrast (white) is additionally shown at the same regular threshold used (P < 0.001) next to the inner patches on both sides. The same overlay is shown with a more liberal (P < 0.01) threshold for the F0 versus ITD contrast in the very right and left panels, respectively. No systematic topographic dissociation between streaming based on f0 and streaming based on ITD can be observed in these overlays.

Figure 6. Time course of fMRI activation in auditory cortex averaged across subjects (n = 12). The time courses for all 3 conditions are based on vertices detected by the ITD + F0 versus NO contrast. Activity from the left (dashed) and right (solid) auditory cortex is overlaid. Note that the activity is more sustained for the streaming conditions (F0 and ITD) compared with the control condition (NO).
temporal gyrus posterior to Heschl’s gyrus plus the planum temporale (posterior ROI). The BOLD response time courses derived from the ITD + F0 versus NO contrast pooled according to the ROIs were analyzed in the same way as above. All ROIs reproduced the effects reported above, with no meaningful differences between them. Therefore, no further details are reported here.

Discussion

Similarity between Streaming Based on Pitch versus Spatial Cues

This study compared activity in auditory cortex during streaming based on a spatial and a pitch cue and found no evidence of topographic stimulus specificity potentially related to streaming. The result supports the hypothesis that the enhancement of neural activity during streaming-related slowing of rate perception is independent of the streaming cue, as proposed by a previous study (Gutschalk, Oxenham, et al. 2007).

In MEG, the main effect during streaming was enhancement of the $P_1 \cdot m$, for both spatial and pitch cues, which is consistent with previous studies using pure tones (Gutschalk et al. 2005; Snyder et al. 2006) and harmonic complex tones with unresolved harmonic structure (Gutschalk, Oxenham, et al. 2007). Both the amount of amplitude enhancement and the location of the dipole sources of the $P_1 \cdot m$ were not different between the pitch and ITD conditions. A similar amplitude enhancement was observed for the subsequent positive wave, the $P_2 \cdot m$, which was more prominent for the pitch cue compared with the ITD cue. This difference between conditions is not necessarily related to streaming, though, because the second positivity comprises a mixture between the $P_2 \cdot m$ evoked by the A tone and the $P_1 \cdot m$ evoked by the first B tone. While the physical characteristics of the A tone were kept identical across conditions, to exclude any effects of physical stimulus differences on the $P_1 \cdot m$, the B-tone $f_0$ was different for the pitch condition. Therefore, the difference of the second positivity may rather be related to these physical differences of the B tones between pitch and ITD conditions.

No streaming-associated modulation of response amplitude was observed for the $N_1 \cdot m$, even though this response was generally present with the paradigm used.

The fMRI results show the enhancement of sustained BOLD signal amplitude for the streaming conditions in comparison to a control, similar to the effects observed by previous fMRI studies of streaming (Gutschalk, Oxenham, et al. 2007; Wilson et al. 2007). This enhancement of sustained BOLD is thought to reflect the change of perceived tone-repetition rate due to streaming similar to the effect produced by a change in physical tone-repetition rate (Harms and Melcher 2002). The change of waveshape from a more phasic to a more sustained pattern was corroborated by quantification with a waveshape index measure (Harms and Melcher 2003). The waveshape index showed a significantly more sustained pattern for the streaming conditions compared with the control, which was more pronounced for the pitch-based streaming. In the present study, the nonstreaming control was limited to a single condition where A and B tones were completely identical. These data therefore cannot exclude that some enhancement of the response could have already occurred for smaller $f_0$ or ITD differences that did not generally produce streaming. However, a previous study with a similar stimulus paradigm (Gutschalk, Oxenham, et al. 2007) found that small $f_0$ differences, which did not produce streaming, also did not produce the same enhancement of MEG and fMRI responses found for conditions that produced streaming.

The present data show that the activity increment extends into anterior as well as posterior areas, comprising primary and nonprimary fields of the auditory cortex (Galaburda and Sanides 1980; Rivier and Clarke 1997). This finding is generally consistent with the ROI-based analysis of Wilson et al. (2007) and Gutschalk, Oxenham, et al. (2007). Moreover, in a direct comparison between spatial and pitch cues, we found no evidence for specificity during either streaming based on a pitch difference nor for streaming based on an ITD difference. The extend of the activity was somewhat larger for the F0 than the ITD condition. This is consistent with the larger $P_2 \cdot m$ component measured by MEG for the F0 condition and most probably unrelated to streaming. Conversely, no vertices were identified within the auditory cortex where the ITD condition produced stronger activation. Within vertices where the pitch condition produced significantly stronger activity than the ITD condition, an ROI analysis revealed the same general pattern as in other ROIs, with both streaming conditions being more sustained than the control and the pitch condition showing a slightly larger amplitude. These vertices rather just show a better SNR but are otherwise not generally different from vertices where no difference between conditions was observed. In summary, it appears that the modulation of neural activity observed during streaming is distributed throughout auditory cortex and does not reflect the particular stimulus cue that the streaming is based on.

Possible Role of Subcortical Processes

One hypothesis to explain the results of the present and previous studies is a relationship between streaming and selective adaptation: It has been suggested that frequency-selective adaptation (or forward suppression) in monkey A1
subserves the stronger separation of 2 different frequency streams at faster rates (Fishman et al. 2001). Moreover, adaptation can develop over longer time intervals and parallels the perceptual buildup of streaming with ambiguous sequences in A1 (Michel et al. 2005). Pressnitzer et al. (2008) showed that similar buildup effects are observed in the cochlear nucleus, suggesting that the observations in A1 might simply reflect a subcortical process whose output is relayed to the auditory cortex. Possibly, some results of the imaging studies (Gutschalk et al. 2005; Wilson et al. 2007) could then be explained by subcortical adaptation as well. Generally, a subcortical model would offer a straightforward explanation of why the cortical enhancement observed with reduced rate perception is similarly observed across multiple areas of the auditory cortex.

However, adaptation in the cochlear nucleus has so far only been shown for pure tones (Pressnitzer et al. 2008), which are readily separated along the tonotopic representation. While there is evidence for early processing of sound features such as pitch (Sayles and Winter 2008) and space (Delgutte et al. 1999) in the brainstem, it is yet unclear whether these representations already subserve the organization into auditory streams at a subcortical level. Moreover, the BOLD response enhancement in auditory cortex associated with streaming is probably related to the change of perceived tone repetition rate (Gutschalk, Oxenham, et al. 2007; Wilson et al. 2007), which was predicted based on the relationship between physical repetition rate and the BOLD response (Harms and Melcher 2002; Wilson et al. 2007). For the rates used here, lower repetition rates were expected to be associated with a higher sustained BOLD level in cortex, but similar effects were not found in the thalamus and the midbrain in that study (Harms and Melcher 2002). Therefore, the presently available data do not support a purely subcortical model of streaming based on selective adaptation.

It would rather appear that the various stages of subcortical feature processing subserve selective adaptation in the primary auditory cortex that would then forward this information as a reference for perceptual organization to the secondary areas of the auditory cortex. Although the exact subcortical and cortical contributions in this process are yet unknown, such a model would comply with most of the observations summarized above and would also match with suggestions of Nelken et al. (2003) for the role of primary auditory cortex as a module to represent auditory streams (or objects).

**Selective Adaptation and Cortical Feature Specificity**

Currently, data showing potential neural correlates for streaming based on nonspectral cues remain confined to the auditory cortex. The previous evidence for streaming based on pitch with spectrally unresolved harmonics (Gutschalk, Oxenham, et al. 2007) is supported and extended to include ITD as a streaming cue by the present study. The existence of amplitude-modulation specificity that could subserve segregation of auditory streams into distinct neuronal populations in the starling forebrain (which may be considered equivalent to the auditory cortex in mammals) was recently demonstrated by Itatani and Klump (2009), although there was little adaptation associated with this activity.

Therefore, one possibility was a purely cortical model for streaming based on more complex features that might rely on feature specificity in nonprimary auditory cortex (Griffiths et al. 1998; Tian et al. 2001; Gutschalk et al. 2002; Patterson et al. 2002; Warren and Griffiths 2003; Penagos et al. 2004; Bendor and Wang 2005; Deouell et al. 2007). Based on the topographical distribution of specificity for pitch and space, however, one would then have expected that the adaptation-related effects were more pronounced (in lateral) Heschl's gyrus when the streaming was based on pitch and more pronounced in the planum temporale when the streaming was based on space (ITD). Selective adaptation reflecting this topography has been previously observed: For example, Warren and Griffiths (2003) showed with fMRI that variation of pitch or space specifically enhances activity in an anterior (pitch) or posterior (space) partition of the auditory cortex. However, their stimulus material was not intended to induce stream segregation but to produce increased complexity within a single stream of tones by variation of pitch and space (using a generic head-related transfer function). In contrast, our pitch and ITD conditions produced the perception of separate streams from 2 distinct sound sources, while the perception of each stream remained a simple monotone sequence (with a changed rate and pattern). Apparently, these different stimulus paradigms recruit distinct systems within the auditory cortex: One system reflects feature specificity with a topographical organization (Warren and Griffiths 2003). The other system may reflect the perceptual organization into streams independently of the feature that the organization is based on.

This dissociation could possibly be transferred to the MEG data, considering that the perceived ISI change associated with streaming was primarily reflected by the $P_{1,m}$ component, which is consistent with previous results (Gutschalk et al. 2005; Snyder et al. 2006; Gutschalk, Oxenham, et al. 2007). On the other hand, the later $N_{1,m}$ showed a strong adaptation within the first few repetitions but was not significantly different between streaming and control conditions. Conversely, studies of pitch perception show specific activity (Krumhholz et al. 2003) and contextual modulation (Gutschalk, Patterson, et al. 2007) of an $N_{1,m}$ component, while the $P_{1,m}$ was unaffected by these stimulus modifications. One possibility would therefore be that the $P_{1,m}$ reflects an earlier distributed system that is potentially related to streaming, whereas the $N_{1,m}$ reflects later processes with longer time constants (Hari et al. 1980; Gutschalk, Patterson, et al. 2007) and is probably related more closely to feature specificity in the auditory belt observed in fMRI (Patterson et al. 2002; Warren and Griffiths 2003; Penagos et al. 2004).

**Possible Role of Top-Down Modulation**

Generally, the rate-related changes of the BOLD response and the $P_{1,m}$ in auditory cortex that covary with streaming might not only be driven by bottom-up processes (as discussed above) but could as well be subject to top-down modulation: Top-down processes in the context of streaming could include feedback from the auditory belt and parabelt areas to the auditory core, especially when the stimulus feature that streaming is based on was more complex (Moore and Gockel 2002). Such a process could potentially modulate activity across the auditory cortex for MEG and fMRI and thereby mask topographic specificity of the activity investigated here. This would, for example, be expected in models where a representation of auditory streams was based on coherence between multiple sites of the auditory cortex (Singer 1999; Elhilali et al. 2003; Penagos et al. 2004).
Top-down modulation of streaming-related processes may not be limited to the cortex but could also involve projections from cortex to the subcortical nuclei (Bajo et al. 2007). Polymodal areas such as the intraparietal sulcus (Cusack 2005) could modulate the representation of streams in the auditory cortex, as well. The contribution of areas outside of the auditory cortex was beyond the scope of the present study, however, whose protocol was focused on the auditory cortex and did not consistently cover other cortical regions in fMRI.

While it appears unlikely at this point that streaming emerges within a single center of the auditory pathway, it is not yet settled which centers are required for a complete model. The auditory cortex is still a main candidate for the convergence of neural representations subserving auditory organization, but subcortical and polymodal areas will also require further evaluation. With respect to the lack of topographically arranged cortical specificity for streaming cues found in the present study, a next step could be to investigate subcortical correlates of streaming based on nonspectral cues, such as pitch and ITD.

**Funding**

Deutsche Forschungsgemeinschaft (GU593/3-1), Bundesministerium für Bildung und Forschung (01EV0712 to A.G.).

**Notes**

We thank Jennifer Melcher for generously providing us the basis functions to estimate the waveshape index. Conflict of Interest: None declared.

**References**


