Cerebral Cortex March 2011;21:698–707
doi:10.1093/cercor/bhq136
Advance Access publication August 4, 2010

Human Auditory Cortex Activity Shows Additive Effects of Spectral and Spatial Cues during Speech Segregation

Yi Du[1,2], Yu He[1], Bernhard Ross[2], Tim Bardouille[2], Xihong Wu[1], Liang Li[1] and Claude Alain[2,3]

[1]Department of Psychology, Speech and Hearing Research Center, Key Laboratory on Machine Perception (Ministry of Education), Peking University, Beijing, China 100871, [2]Rotman Research Institute, Baycrest Centre for Geriatric Care, Toronto, Ontario, Canada M6A 2E1 and [3]Department of Psychology, University of Toronto, Ontario, Canada M8V 2S4

Address correspondence to Claude Alain, Rotman Research Institute, Baycrest Centre for Geriatric Care, 3560 Bathurst Street, Toronto, Ontario M6A 2E1 Canada. Email: calain@rotman-baycrest.on.ca.

In noisy social gatherings, listeners perceptually integrate sounds originating from one person's voice (e.g., fundamental frequency (f_0) and harmonics) at a particular location and segregate these from concurrent sounds of other talkers. Though increasing the spectral or the spatial distance between talkers promotes speech segregation, synergetic effects of spatial and spectral distances are less well understood. We studied how spectral and/or spatial distances between 2 simultaneously presented steady-state vowels contribute to perception and activation in auditory cortex using magnetoencephalography. Participants were more accurate in identifying both vowels when they differed in f_0 and location than when they differed in a single cue only or when they shared the same f_0 and location. The combined effect of f_0 and location differences closely matched the sum of single effects. The improvement in concurrent vowel identification coincided with an object-related negativity that peaked at about 140 ms after vowel onset. The combined effect of f_0 and location closely matched the sum of the single effects even though vowels with different f_0, location, or both generated different time courses of neuromagnetic activity. We propose that during auditory scene analysis, acoustic differences among the various sources are combined linearly to increase the perceptual distance between the co-occurring sound objects.

Keywords: attention, MEG, scene analysis, streaming, speech

Introduction

Auditory scene analysis is a dynamic process that allows listeners to identify and localize concomitant sound sources (i.e., auditory objects) in the environment. In social gathering, this entails grouping together those sound elements coming from one source (i.e., one speaker) and segregating those arising from other sources (another speaker). Early studies have shown that paying attention to what a person is saying in the presence of other talkers is facilitated by increasing spectral (e.g., voice) or spatial (e.g., ear) differences between sound sources (Spieth et al. 1954; Treisman 1964). Subsequent studies have also shown substantial benefit from spectral (e.g., Chalikia and Bregman 1989; Assmann and Summerfield 1990, 1994) or spatial (Shackleton and Meddis 1992; Drennan et al. 2003) differences in identifying 2 different vowels presented simultaneously. In the studies mentioned above, the effects of spectral and spatial differences on concurrent speech separation and identification were investigated separately. Yet, in complex listening situations, the various sound objects composing the auditory scenes often differ in terms of their spectral signature as well as the spatial direction of sound origin. Hence, the separation and identification of concurrent speech sounds may be enhanced by a process that integrates the differences in fundamental frequency (f_0) and spatial location. For instance, both the effect of f_0 and spatial separation may contribute to concurrent speech separation and identification in an additive or superadditive manner. Alternatively, separation of concurrent speech sounds may be driven primarily by the most salient cue (e.g., f_0) independently of the other cue (i.e., location). For instance, separate decision processes could be initiated simultaneously for f_0 and location with performance and brain activity being driven by the quicker process (horse-race model, Mordkoff and Yantis 1991). Studying how listeners perceptually organize concurrent speech sounds that differ in f_0 and/or location may reveal important processing principles related to concurrent sound perception under ecologically valid circumstances.

Animal studies have revealed linear and nonlinear interactions between different perceptual features in primary and nonprimary auditory cortices (Machens et al. 2004; Ahmed et al. 2006; Bizley et al. 2009). In humans, evidence for linearity in auditory cortex comes from studies employing the oddball paradigm, where the mismatch negativity (MMN) wave elicited by sounds that deviate from the standards along 2 perceptual dimensions equals the sum of MMNs elicited by each deviant dimension presented alone. Such additivity has been observed for several dimension combinations including frequency and location (Schröger 1995; Takegata and Morotomi 1999; Paavilainen et al. 2001), frequency and stimulus onset asynchrony (Levanen et al. 1993) and, frequency and duration (Levanen et al. 1993; Wolff and Schröger 2001). The observed additivity for effects of frequency and location suggests that these stimulus dimensions are initially processed independently from each another. During auditory selective attention, this initial additivity for frequency and location is followed by nonlinear interactions with attention-related neural activity to stimuli defined by a combination of features (e.g., pitch and location) differing from the sum of the attention effects elicited by each feature alone (Woods and Alain 1993, 2001; Woods et al. 1994).

Together, the studies reviewed above showed additivity in auditory cortex for simple tones. However, as currently understood, the additivity principle may not adequately explain the perceptual grouping of speech sounds (Remez et al. 1994, 2008) because speech has acoustic properties that are diverse and rapidly changing. Furthermore, speech is a highly familiar stimulus, and as such our auditory system has had the opportunity to learn about speech-specific properties (e.g.,

fundamental frequency and formant transitions) that may assist in the successful perceptual grouping of speech stimuli (Rossi-Katz and Arehart 2009). At present, the mechanisms involved in parsing concurrent speech are unclear but may involve linear and/or nonlinear integration of acoustic elements to increase the perceptual distance between co-occurring sound objects.

We studied how the spectral and spatial differences between 2 simultaneously presented vowels contribute to auditory cortical activity measured with magnetoencephalography (MEG). First, we established the optimal location separation to yield improvement in vowel identification. Then, we compared neuromagnetic activity when both vowels shared the same f_0 and location with conditions where they differed in f_0, location, or both. Previous work showed an early object-related negativity (ORN, ~150 ms) and a later (~250 ms) evoked response that reflected the automatic registration of f_0 differences and higher-order decision-making mechanisms, respectively (Alain et al. 2005). We anticipated behavioral improvement when the 2 vowels differed in f_0 and location that would correlate with changes in neuromagnetic activity from the auditory cortex. If concurrent speech separation relies on a process that combines information about f_0 and location, then accuracy and ORN amplitude will show additive effects. On the other hand, if concurrent speech separation and identification are determined primarily by the most salient cue (e.g., f_0) independently of the other cue (i.e., location), then this would argue in favor of a horse-race model (e.g., Mordkoff and Yantis 1991) with accuracy and ORN being driven by the quicker process.

Material and Methods

Participants

A total of 27 participants provided written informed consent to participate in the study. In Experiment 1, there were 8 women and 4 men aged between 20 and 34 years (mean age = 25 ± 3.7 years). In Experiment 2, 6 women and 6 men (none of whom participated in Experiment 1) aged between 21 and 33 years (mean age = 26 ± 3.8 years) were included in the analysis. Three participants were excluded from the analysis due to excessive head motion during MEG recording. All participants, except one in Experiment 2, were right-handed and all had pure-tone thresholds within normal limits for octave frequencies ranging from 250 to 4000 Hz (both ears). For all participants, English was the first language. Ethical approval and informed consent were obtained according to the guidelines set out by Baycrest Centre and the University of Toronto.

Stimuli and Task

In both experiments, stimuli were 4 synthetic steady-state American English vowels: /a/, /ɜ/, /i/, /u/ (Assmann and Summerfield 1994). Each vowel was 200 ms in duration with f_0 and formant frequencies held constant. Stimuli were converted to analog forms using a TDT RP-2 real-time processor (Tucker Davis Technologies). The analog outputs were fed into a headphone driver (TDT HB-7) and presented binaurally at 75 dB sound pressure level through Etymotic ER3A insert earphones connected with 1.5 m of reflectionless plastic tubing. The frequency response of the ER3A insert earphones was flat from 80 to 2000 Hz within ±5 dB. Beyond 2000 Hz, the frequency response decreased gradually. Sound locations' differences were induced by applying a head-related transfer function (HRTF) to the vowels with the coefficients taken from the TDT library (for a detailed description and behavioral validation of the HRTF coefficient used, see Wightman and Kistler 1989a, 1989b; Wenzel et al. 1993).

Prior to Experiments 1 and 2, each vowel was presented individually (16 trials, 4 vowels by 2 f_0 levels), and participants identified the vowels by pressing the corresponding keys. All participants identified the single vowels with an accuracy of 95% or better. In order to find the most suitable HRTF coefficients, each participant completed a behavioral calibration task which required them to identify (by pointing) the locations of several vowels. The stimuli were presented at various locations using a variety of HRTF coefficients that best suit the head size. The coefficients that resulted in the most accurate responses were then determined and used for the remainder of the experiment.

In Experiment 1, each vowel pair contained 2 vowels with the same f_0 (either at 100 or at 106 Hz) but coming from either the same (both straight ahead, 0°) or different locations (one from the left, the other from the right, and both were 15°, 30°, 45°, 60°, or 75° away from the midline). This resulted in 6 levels of location difference between the 2 vowels: 0°, 30°, 60°, 90°, 120°, or 150°. All vowel pairs were presented in the horizontal plane, and the 2 vowels' locations changed from trial to trial with equal probability for each level of spatial separation. Vowel pairs were presented in random order in blocks of 72 trials. Each participant completed 6 blocks of trials.

In Experiment 2, there were 4 different stimulus types created by the orthogonal combination of f_0 and location differences. That is, the 2 vowels could have either the same (100 or 106 Hz) or different f_0 (one at 100 Hz, the other at 106 Hz, i.e., 1-semitone difference ($\triangle f_0$)), and they could either come from the same (0°) or different azimuth locations (one from 45° to the left, the other from 45° to the right, i.e., 90° location difference (△location)). These 4 stimulus types were labeled as follows: same f_0-same location (SFSL), different f_0-same location (DFSL), same f_0-different location (SFDL), and different f_0-different location (DFDL). Each stimulus type occurred with the same probability (25%). The 4 different stimulus types were presented in random order in blocks of 144 trials, and 4 blocks of trials were presented to each participant.

The f_0 separation and location separation between the 2 vowels was set at one semitone and 90°, respectively. The choice of these parameters was based on prior research showing that participant's accuracy reached an asymptote when the 2 different vowels are separated by one semitone (Assmann and Summerfield 1990; Summerfield and Assmann 1991; Alain et al. 2005) and when they are separated by 90°–120° (see Result's section Experiment 1).

In both experiments, listeners were instructed to identify both vowels in the pair. They registered their responses by sequentially pressing 1 of 4 keys on the keyboard, marked "AH," "ER," "EE," and "OO" for the vowels /a/, /ɜ/, /i/, /u/, respectively. Participants were told that 2 different vowels would always be presented in each trial. The interval between the participant's response and the next trial was 1500 ms. No feedback was provided after each response.

Data Acquisition

In Experiment 1, behavioral data were only collected. The proportion of trials in which both vowels were correctly identified as a function of the difference in location was subjected to a within-subject repeated measures analysis of variance (ANOVA).

In Experiment 2, the neuromagnetic brain activity was recorded in a magnetically shielded room using a 151-channel whole-head neuromagnetometer (VSM Medtech). Participants were positioned in the upright seating position with their head resting in the helmet-shaped scanner. Head localization coils were placed on the nasion, and on left and right preauricular points prior to scanning for coregistration of neuromagnetic data with anatomical magnetic resonance images (MRIs). Realistic estimates of the participants' head shapes were also acquired with a 3D digitization system (Fastrak; Polhemus). The neuromagnetic activity was collected for 4 blocks, each block lasting about 9 min, with a sampling rate of 625 Hz and low-pass filtered at 200 Hz.

Data Analysis

We performed 2 complementary types of data analysis. First, for each participant, we modeled the grand average auditory evoked fields (AEFs) with single dipoles in left and right auditory cortices and calculated the waveforms source strength. This allowed us to study the sequence of positive and negative waves in the responses as occurs in

the more conventional analysis of selected electrodes in electroencephalography or sensors in MEG. Second, we performed a model-free analysis of source activity in order to identify the brain areas involved in processing f_0 and location differences. However, in this paper, we will mainly focus on the neural interaction between f_0 and location differences during the early stages of processing and encoding in sensory memory, therefore, only neural activities in the bilateral superior temporal gyrus are presented.

Dipole Source Analysis

We used BESA software version 5.2 (Brain Electrical Source Analysis, MEGIS Software GmbH) for averaging and dipole source modeling. The analysis epoch included 200 ms of prestimulus activity and 800 ms of poststimulus activity. The artifact rejection threshold was adjusted for each participant such that about 90% of trials were included in the average. The threshold for rejecting single epochs of MEG data ranged from 2300 to 4420 fT. AEFs were averaged separately for each stimulus type. For all stimulus conditions, the number of trials included in the averages varied between 124 and 141 trials (M = 133). For each participant, we also computed the grand average of AEFs that comprised all stimulus conditions. Before dipole source modeling, the averaged data were low-pass filtered at 20 Hz (12 dB/octave; zero phase).

A dipole source model including a left and a right dipole in the temporal lobes was used as a data reduction method. For each participant, the dipole source localization was performed on the data grand averaged across stimulus types. First, the dipoles were seeded in the temporal lobe near Heschl's gyrus and then the location and orientation of each dipole was fit to account for a 40-ms interval centered on the peak of the N1m wave. We chose to model the N1m wave because it was the largest and most reliable deflection from the AEF elicited by the double-vowel stimuli. The group mean residual variance for the source model was 18.0% (standard error = 1.06%). Following this, the location and orientation of the dipoles were kept constant and the source waveforms were extracted for each stimulus condition in each participant. Peak amplitude and latency were determined as the largest positivity or negativity in the individual source waveforms during a specific interval. The measurement intervals were 30–70 ms, 70–170 ms, and 160–260 ms relative to sound onset, for P1m, N1m, and P2m, respectively.

For the behavioral data in Experiment 2, a repeated measures ANOVA was conducted on the proportion of trials in which both vowels were correctly identified with f_0 (same, different) and location (same, different) as the within-subjects factors. For the MEG data, the repeated measures ANOVAs were performed on P1m, N1m, and P2m peak amplitudes and latencies and mean amplitudes (nAm) of source waveforms over selected latency regions with f_0 (same, different), location (same, different) and hemisphere (left, right) as the within-subjects factors.

Event-Related Beamformer Analysis

The synthetic aperture magnetometry (SAM) minimum-variance beamformer algorithm (Robinson and Rose 1992; Van Veen et al. 1997; Robinson and Vrba 1998) was used as a spatial filter to estimate the time course of source activity on a lattice of 5 mm spacing across the whole brain volume in the 0.3–20 Hz frequency range. A multiple sphere head model was used for the beamformer analysis in which a single sphere was fit to the digitized head shape for each MEG sensor. Waveforms of averaged source activity across all trials for each stimulus type were calculated following the event-related SAM (ER-SAM) approach (Robinson 2004; Cheyne et al. 2006). The ER-SAM procedure results in noise-normalized estimates of source power termed "pseudo Z" values (Robinson and Vrba 1998; Chau et al. 2004). Volumetric maps of group mean pseudo Z values as a function of time for each stimulus type were overlaid on the anatomical image of a template brain (colin27, Montreal Neurological Institute) (Holmes et al. 1998) and were visualized with AFNI software (National Institute of Mental Health) (Cox 1996). For source waveforms at the voxel showing maximum modulation of activity within Heschl's gyrus, repeated measures ANOVAs were performed on N1m (80–120 ms) peak source amplitudes and latencies and mean source amplitudes over selected latency regions to test the effects of f_0 and location differences and their interactions.

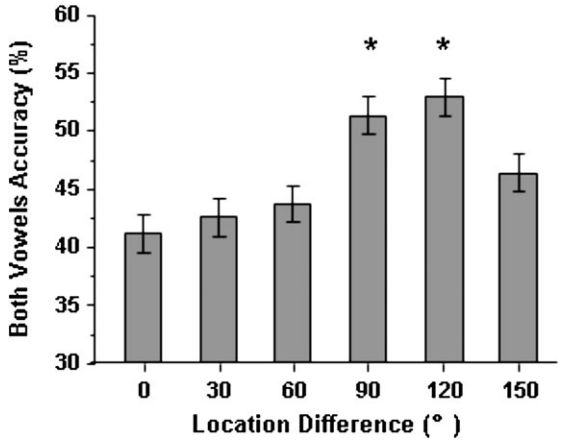

Figure 1. Behavioral performance in Experiment 1. Proportion of trials in which both vowels were correctly identified is plotted as a function of the difference in location between the 2 vowels. The error bars (±standard error of the mean) indicate the within-subject variability at each location separation. * $P < 0.05$ relative to all other conditions.

Results

Experiment 1

Figure 1 shows the group mean proportion of trials in which both vowels were correctly identified as a function of △location. Participants performed well above chance (i.e., chance level for identifying both vowels was 17%), even when the 2 vowels shared the same location. The main effect of △location was significant, $F_{5,55} = 17.95$, $P < 0.001$, with participants being more accurate when the 2 vowels were separated by 90° and 120° compared to when they were presented at the same location or when they were separated by only 60° ($P < 0.05$ in all cases). There was no significant improvement from 90° to 120° separation. However, vowel identification significantly decreased from 120° to 150° separation ($P < 0.01$). There was no significant difference in performance when both vowels were presented at the same location or when they were separated by 30°, 60°, or 150°. These results show that a 90–120° separation in location was optimal to effectively improve concurrent vowel segregation and identification.

Experiment 2

Accuracy

The effects of f_0 and/or spatial separation on accuracy are illustrated in Figure 2*A*. The proportion of trials in which both vowels were correctly identified improved with increasing $\triangle f_0$ and/or △location between the 2 vowels. The main effects of $\triangle f_0$ and △location were significant, $F_{1,11} = 24.57$ and 44.54, respectively, $P < 0.001$ in both cases. The interaction between f_0 and location was not significant, $F_{1,11} < 1$.

Figure 2*B* compares the benefits of having both $\triangle f_0$ and △location against the linear sum of the main effects of $\triangle f_0$ and △location. A repeated measures ANOVA indicates that the benefit in identification rate differed across stimulus types,

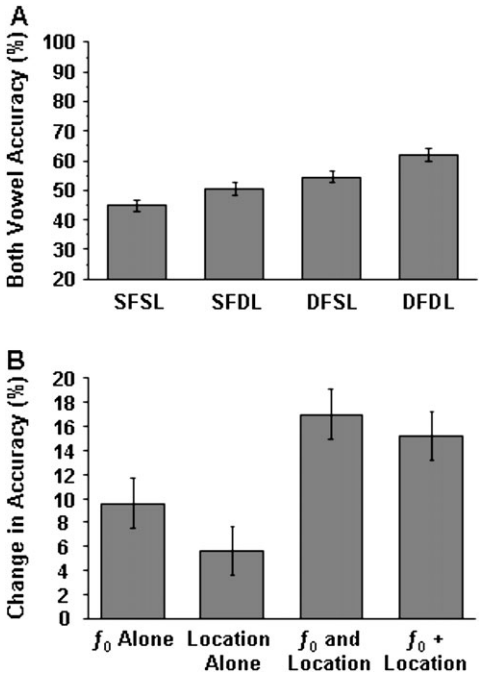

Figure 2. Behavioral performance and behavioral benefit in Experiment 2. (*A*) Proportion of trials in which both vowels were correctly identified under 4 stimulus types: SFSL, SFDL, DFSL, and DFDL. (*B*) Changes in accuracy of identification of both vowels (compared with performance when 2 vowels shared same f_0 and location) for stimulus conditions with only $\triangle f_0$ (f_0 alone), $\triangle$location (location alone), and both $\triangle f_0$ and $\triangle$location (f_0 and location). A linear sum of change by $\triangle f_0$ alone and that by $\triangle$location alone is shown also (f_0 + location). The error bars (±standard error of the mean) indicate the within-subject variability for each condition.

$F_{3,33}$ = 13.39, P < 0.001. Pairwise comparisons show that the benefit of having both $\triangle f_0$ and $\triangle$location was significantly larger than the main effect of $\triangle f_0$ or $\triangle$location alone (P < 0.05 in both cases). The effect sizes for $\triangle f_0$ and $\triangle$location alone were not significantly different from each other. The benefit of having both $\triangle f_0$ and $\triangle$location did not significantly differ from the linear sum of $\triangle f_0$ and $\triangle$location alone.

Dipole Source Waveforms

Figure 3*A* overlays the time courses of the AEFs recorded with all MEG sensors in response to the double-vowel stimuli in one representative participant. The AEFs comprise an initial P1m peak at 42 ms followed by the larger N1m at 110 ms and long-lasting negativity (i.e., same polarity as N1m) with maximum around 400 ms. This long-lasting negativity overlaid the P2m responses at 205 ms. The magnetic field topography at the latency of the N1m is consistent with bilateral sources in the superior temporal gyri (Fig. 3*B*). The dipole locations for the N1m for that particular participant as well as the group mean locations for the N1m are shown in Figure 3*C*. The group mean N1m dipole locations were in the left (Talairach coordinates: –47, –27, 7) and right (46, –23, 5) superior temporal gyri posterior to Heschl's gyrus. The dipole in the right hemisphere was more anterior (t_{11} = 4.32, P < 0.01), and inferior (t_{11} = 2.24, P < 0.05) to the dipole in the left hemisphere. There was no difference along the medial–lateral axis.

Figure 4*A* shows the group mean dipole source waveforms as a function of $\triangle f_0$ and/or $\triangle$location. There was no difference in P1m latency or amplitude as a function of stimulus type. The first reliable effect of stimulus condition emerged during the N1m interval. The N1m peak latency was earlier when the 2 vowels shared the same f_0 as compared to when they differed in f_0, $F_{1,11}$ = 12.00, P < 0.01. No latency difference based on spatial separation was found for the N1m, $F_{1,11}$ < 1. An ANOVA on the N1m peak amplitude yielded a main effect of $\triangle$location, $F_{1,11}$ = 28.95, P < 0.001. The main effect of $\triangle f_0$ was not significant nor was the interaction between $\triangle f_0$ and $\triangle$location. There was no hemispheric difference for the N1m amplitude nor was the interaction between hemisphere and stimulus condition significant.

The P2m latency was not significantly affected by either f_0 or location but was slightly longer in the right hemisphere, $F_{1,11}$ = 4.84, P = 0.05. The P2m was smaller when the 2 vowels were presented at different locations than when they were presented at the same location, $F_{1,11}$ = 27.57, P < 0.001. The main effect of f_0 on P2m amplitude was not significant, $F_{1,11}$ = 2.96, P = 0.113. However, there was a significant interaction between f_0 and hemisphere, $F_{1,11}$ = 5.40, P < 0.05. Separate ANOVAs for the left and right hemispheres revealed a main effect of f_0 only in the right hemisphere, $F_{1,11}$ = 9.57, P = .01.

Object-Related Negativity

The neural activity associated with the effects of $\triangle f_0$ and/or $\triangle$location on concurrent vowel identification is best illustrated by subtracting dipole source waveforms when both vowels shared the same f_0 and location from those acquired when the 2 vowels differed in f_0 only, location only, or both f_0 and location. This subtraction procedure revealed an early positive peak (~70 ms), the ORN (~120 ms for stimuli with $\triangle$location alone and both $\triangle f_0$ and $\triangle$location, ~160 ms for stimuli with $\triangle f_0$ alone), and a second negative peak (~230 ms) hereafter referred to as N2b (Fig. 4*B*).

Consistent with previous work using the mistuned harmonic paradigm (Alain and McDonald 2007), difference in f_0 between the 2 vowels yielded a reliable increase in source waveform mean amplitude during the 50- to 90-ms intervals, $F_{1,11}$ = 13.19, P < 0.01. The main effect of $\triangle$location was not significant nor was the interaction between $\triangle f_0$ and $\triangle$location. This early registration of $\triangle f_0$ was followed by a main effect of $\triangle$location between 100 and 140 ms after double-vowel onset, $F_{1,11}$ = 23.98, P < 0.001. During this interval, the main effects of $\triangle f_0$ and hemisphere were not significant nor were any of the interactions among $\triangle f_0$, $\triangle$location, or hemisphere. During the 140- to 180-ms intervals, $\triangle f_0$ and $\triangle$location yielded a pronounced ORN, $F_{1,11}$ = 5.63 and 13.23, respectively, P < 0.05 in both cases. The main effect of hemisphere was not significant nor were any of the interactions among $\triangle f_0$, $\triangle$location, and hemisphere.

For the N2b, the ANOVA on the mean amplitude for the 210- to 250-ms intervals yielded a main effect of $\triangle$location, $F_{1,11}$ = 51.38, P < 0.001, and a main effect of hemisphere, $F_{1,11}$ = 6.19, P < 0.05. The main effect of $\triangle f_0$ was not significant. However, there was a significant interaction between $\triangle f_0$ and hemisphere, $F_{1,11}$ = 5.75, P < 0.05. A separate ANOVA for the right hemisphere reveals significant main effects of $\triangle f_0$, $F_{1,11}$ = 5.06, P < 0.05, and location, $F_{1,11}$ = 27.56, P < 0.001 but not for the $\triangle f_0$ × $\triangle$location interaction.

With Figure 4*C*, we compared the difference waveforms for the effect of the 2 simultaneous cues with the sum of difference waveforms observed for the individual cues (The additivity-logic used here is similar to that used for MMN studies and consisted of combining ORNs [difference waves]

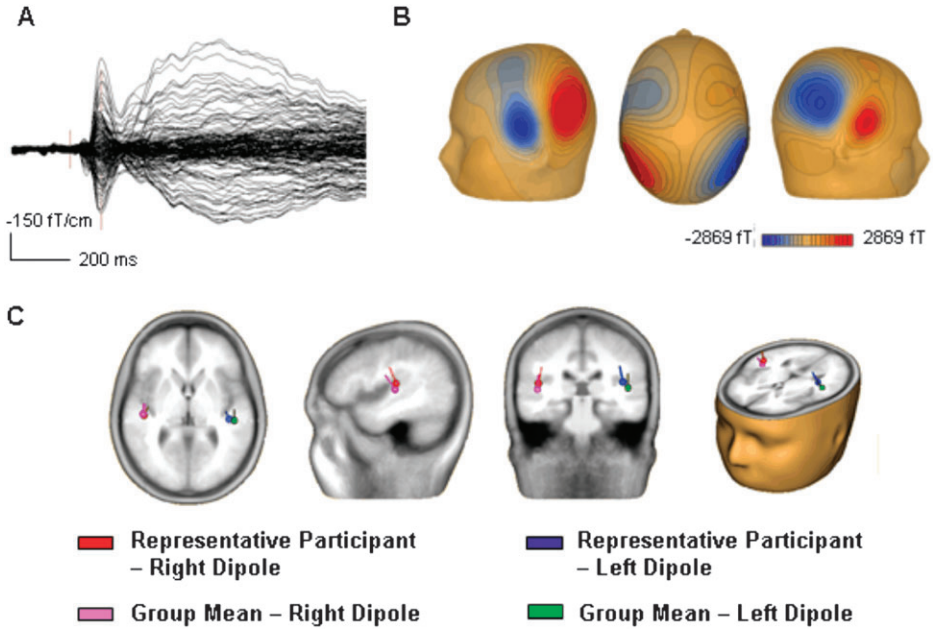

Figure 3. (*A*) AEFs elicited by the double-vowel stimuli in one representative participant. (*B*) The contour maps for the N1m for that particular participant. (*C*) The dipole location for the N1m for that particular participant as well as the group mean location overlaid in MRI template from BESA (5.2).

from different conditions. This approach eliminated unspecific activity that may have contaminated early sensory-evoked responses [Gondan and Roder 2006].). Most interestingly, both graphs match very well as indicated by the difference between both (red line in Fig. 4*C*) showing no prominent deviation from zero. Although there was an early (50–90 ms) disparity between the 2, a repeated measures ANOVA on the mean amplitudes during this interval did not yield a significant difference, $F_{1,11} = 3.54$, $P = 0.087$. Also, there was no significant difference during 140- to 180-ms intervals or 210- to 250-ms intervals, $F < 1$ in both cases. In addition to testing for the null hypothesis, we used the city-block distance (CBD) method (Schröger 1998) to examine whether the ORN elicited by both f_0 and location was similar to the sum of ORNs (140–180 ms) elicited by f_0 only and location only. This analysis revealed significant similarity between the sum of the single and the ORN elicited by both f_0 and location separation. For the source waveforms from the right hemisphere, the sum of absolute difference between the 2 conditions (i.e., the CBD) was 53.03 nAm, $P < 0.01$. For the left hemisphere, the CBD was 42.46 nAm, $P < 0.01$. Together, these results suggest that auditory cortical activities elicited by $\triangle f_0$ and $\triangle$location might be linearly added when both cues are available and provide further support for the notion that frequency and location are independently represented in the auditory cortex.

Brain–Behavior Correlations

Figure 5 illustrates the relation between individual changes in ORN amplitude and the listeners' performance in identification of both vowels. The ORN elicited by $\triangle f_0$ and/or $\triangle$location appeared more negative when higher accuracy was attained in identifying both vowels. To quantify this relationship, the Pearson correlation coefficient was calculated for each condition. For the right hemisphere, this analysis reveals significant correlations between the ORN amplitude and the improvement in identification for $\triangle f_0$ alone ($r = -0.61$, $P < 0.05$), $\triangle$location alone ($r = -0.60$, $P < 0.05$) but not for both $\triangle f_0$ and $\triangle$location ($r = -0.20$, $P > 0.05$). After 3 participants who had very large ORN amplitude were excluded from analysis, the correlation coefficient for both $\triangle f_0$ and $\triangle$location became –0.65 ($P = 0.06$). In comparison, no significant correlation was found between the ORN amplitude from the left hemisphere and the change in participants' performance for each condition, the Pearson correlation coefficients were 0.21, –0.51, and –0.16 for $\triangle f_0$ alone, $\triangle$location alone, and both $\triangle f_0$ and $\triangle$location, respectively.

ER-SAM Source Activity

During the N1m interval, the results from the beamformer spatial filter revealed bilateral sources in auditory cortices along the superior temporal plane. Figure 6 (panel A) shows ER-SAM maps of activation overlaid on a structural MR template from AFNI software for the SFSL stimulus condition. The peak activation in the left (Talairach coordinates: –51, –26, 12) and right (51, –22, 12) auditory cortices closely match those observed for the spatiotemporal dipole source locations.

The time courses of source activation for SFSL condition showed a dominant peak in the latency range of the N1m response with a mean latency of 98 and 101 ms in the left and right auditory cortex, respectively. The repeated measures ANOVA on the N1m peak source amplitudes for 4 stimulus conditions (Fig. 6, panel B) revealed a main effect of $\triangle$location, $F_{1,10} = 7.62$, $P = 0.02$. The main effect of $\triangle f_0$ was not significant nor was the interaction between $\triangle f_0$ and $\triangle$location. There was no hemispheric difference for the N1m peak source amplitudes nor was the interaction between hemisphere and stimulus condition significant. The ANOVA on N1m source latencies did not yield any significant effects of stimulus type or hemisphere.

The differences in source waveforms between conditions were also calculated to illustrate the neural activity associated with the effects of $\triangle f_0$ and/or $\triangle$location on concurrent vowel identification. To this end, we subtracted source waveforms

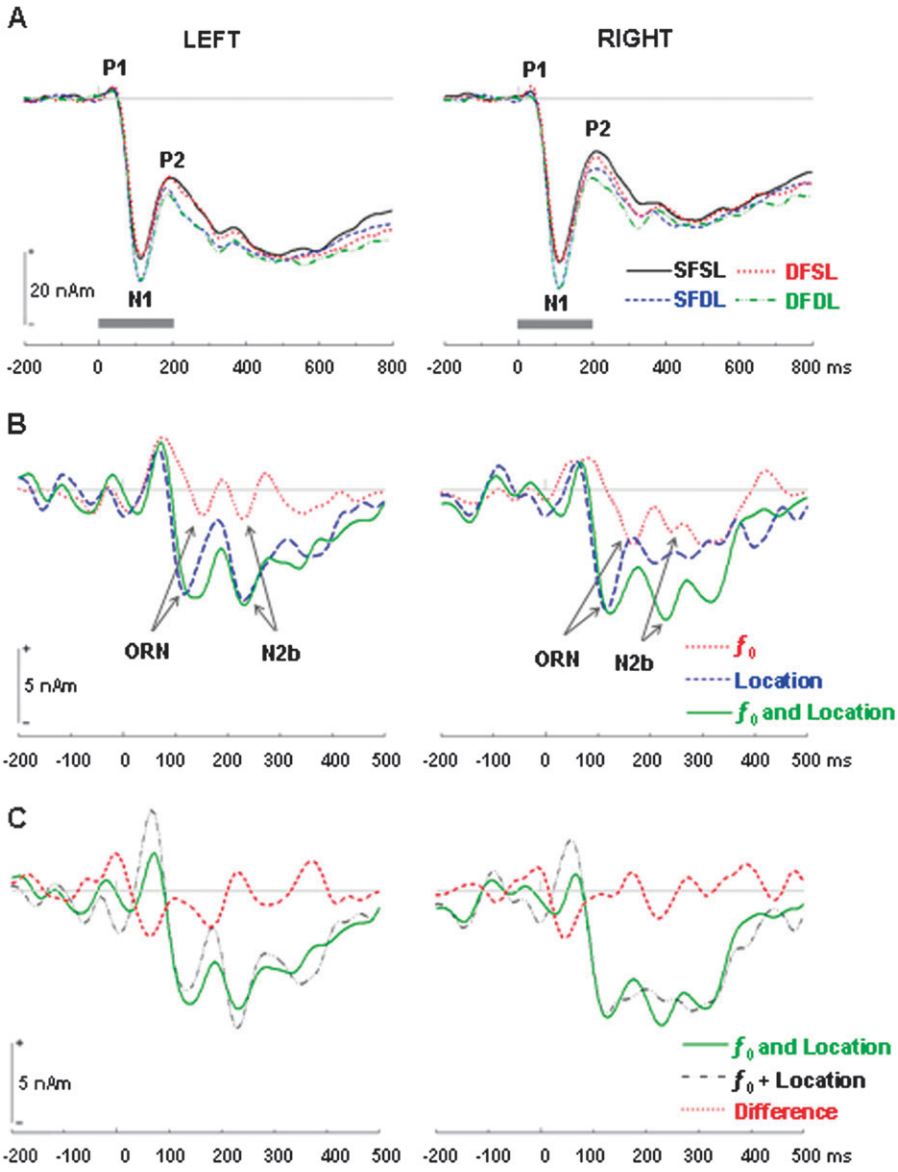

Figure 4. Group mean dipole source waveforms. (*A*) Group mean source waveforms for AEFs for 4 stimulus conditions. The gray rectangles represent the duration of the double-vowel stimulus. (*B*) The differences in source waveforms between stimuli with same f_0 and same location and stimuli with $\triangle f_0$ and/or $\triangle$location. (*C*) Comparison between difference waveforms for both $\triangle f_0$ and $\triangle$location (f_0 and location) and a linear sum of difference waveforms for $\triangle f_0$ alone and that for $\triangle$location alone (f_0 + location).

when both vowels shared the same f_0 and location from those acquired when the 2 vowels differed in f_0 alone, location alone, or both f_0 and location (Fig. 6, panel B). This subtraction procedure reveals a positive peak in the N1m period (~120 ms) for $\triangle$location (SFDL-SFSL, in Fig. 6), a negative peak in N1m period (~104 ms), and a late positive peak (~176 ms) for $\triangle f_0$ (DFSL-SFSL). Again, the difference in source waveforms for both $\triangle f_0$ and $\triangle$location (DFDL-SFSL) closely matched the linear sum of the difference in source waveforms for $\triangle f_0$ alone and that for $\triangle$location alone [(SFDL-SFSL) + (DFSL-SFSL)] in both hemispheres. The repeated measures ANOVAs on the mean amplitudes of these 2 differences in source waveforms during 104- to 144-ms intervals and 152- to 192-ms intervals did not yield a significant difference regardless of the hemisphere, $F < 2$ in both cases. This result again confirms that auditory cortical activities elicited by $\triangle f_0$ and $\triangle$location are linearly added when both cues are available.

Discussion

In most everyday situations, co-occurring sound sources (i.e., auditory objects) usually differ in their spectro-temporal signature as well as in their actual location in the environment. The present study shows that differences in f_0 and/or location between the 2 vowels contribute to speech separation and identification. The effects of $\triangle f_0$ on behavioral performance and neuromagnetic activity were consistent with prior behavioral research (Chalikia and Bregman 1989; Assmann and Summerfield 1990) as well as a prior event-related potential (ERP) study using a greater range of f_0 separation (Alain et al. 2005). The effects of $\triangle$location on concurrent vowel perception were also consistent with findings from a prior behavioral study which also found that ~90° separation in free field or simulated auditory location using HRTF yields the greatest improvement in performance (Drennan et al. 2003). In the present study, performance in identifying both

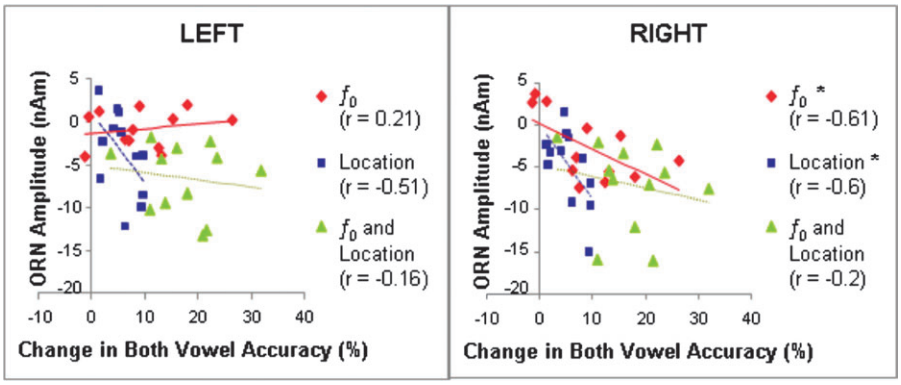

Figure 5. Brain–behavior correlation. Individual changes in source waveform amplitude during the ORN interval (140–180 ms) are plotted against the listeners' changes in accuracy of identification of both vowels for stimuli with $\triangle f_0$ alone (f_0), $\triangle$location alone (location), and both $\triangle f_0$ and $\triangle$location (f_0 and location). $^*P < 0.05$.

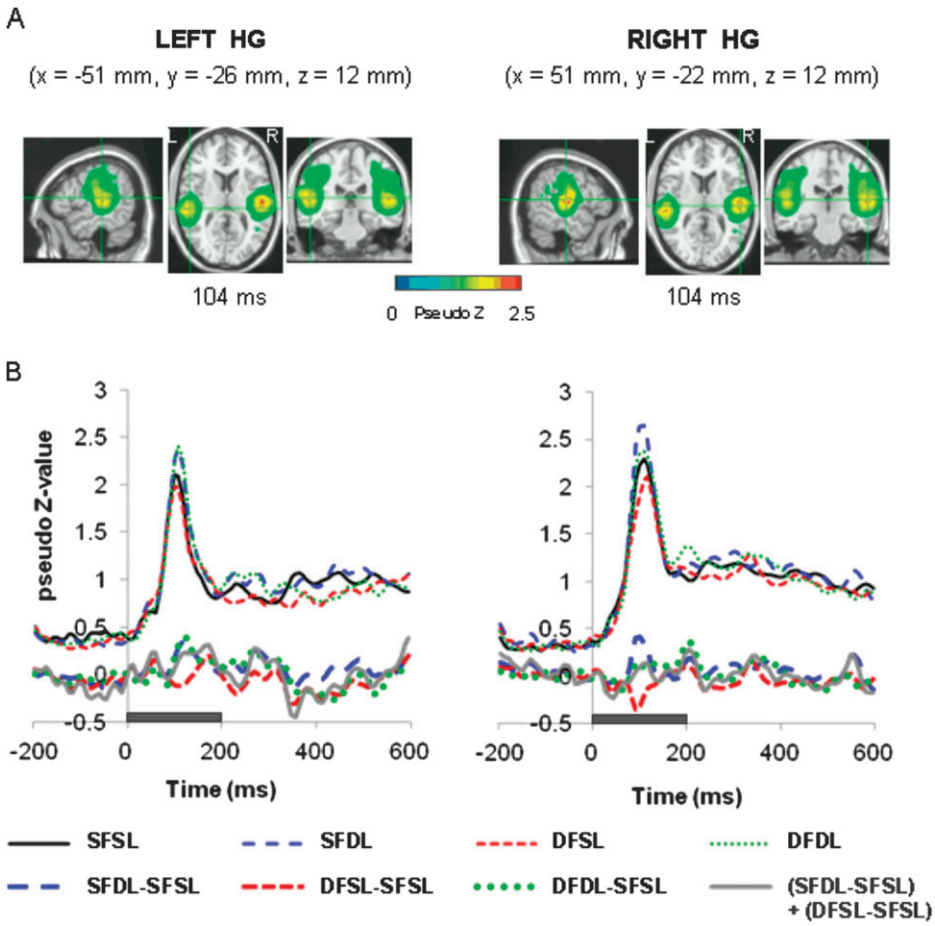

Figure 6. ER-SAM maps and source waveforms in left and right Heschl's gyri. (*A*) Thresholded group-mean ER-SAM maps for SFSL condition at N1m latency of 104 ms. (*B*) Time courses of source activities associated with 4 stimulus conditions (SFSL, SFDL, DFSL, and DFDL) and 4 differences between conditions [SFDL-SFSL, DFSL-SFSL, DFDL-SFSL, (SFDL-SFSL) + (DFSL-SFSL)]. Pseudo *Z* values represent the ratio of signal-to-noise power of the evoked response. Note that in both hemispheres, the difference in source waveforms for both $\triangle f_0$ and $\triangle$location (DFDL-SFSL) closely matches the linear sum of $\triangle f_0$ and $\triangle$location alone [(SFDL-SFSL) + (DFSL-SFSL)].

vowels plateau for 90° and 120° and worsened when the spatial separation between the 2 was further increased to 150°. This biphasic pattern of response was unexpected and could reflect a cost in spreading attention between 2 disparate locations. It could also be related to impoverished spatial resolution for sounds presented further from the midline location. Additional research is needed to better understand the mechanisms that underlie this biphasic pattern of response.

In the present study, the effect of $\triangle f_0$ and $\triangle$location on concurrent vowel identification was comparable. This is not surprising given that the magnitude of $\triangle f_0$ and $\triangle$location were chosen because they yielded asymptotic performance in most listeners. More importantly, performance improved significantly when both $\triangle f_0$ and $\triangle$location were present. The combined effect of $\triangle f_0$ and $\triangle$location equaled the sum of each $\triangle f_0$ and $\triangle$location. The additive effect of $\triangle f_0$ and

△location suggests that auditory scene analysis may involve processes that "sample" and use multiple cues rather than relying on one particular cue in segregating the incoming acoustic data. Note that the additivity observed in the present study may be specific to cases where cues promoting segregation are equally salient.

Both $\triangle f_0$ and △location alone elicited an ORN during 120- to 160-ms intervals following sound onset in auditory cortices. The ORN amplitude recorded from the right hemisphere correlated with listeners' identification accuracy of both vowels. This is consistent with a previous study manipulating only $\triangle f_0$ between concurrent vowels (Alain et al. 2005). In the present study, there was no significant correlation between the ORN recorded from the left hemisphere and accuracy. The reasons for this lack of correlation could be related to smaller and more variable ORN amplitude in the left than in the right hemisphere.

It is suggested that the ORN may index the automatic detection of multiple concurrent sound objects (Alain 2007). In this case, either the separation in f_0 or spatial location of the 2 vowels may promote the activation of 2 distinct neural traces, one for each vowel constituent that eventually facilitates the identification of both vowels. Interestingly, the ORN for △location was 40 ms earlier and much larger in amplitude than that for $\triangle f_0$, despite comparable performance in identifying both vowels. The 2 vowels presented at different locations likely activate distinct perceptual channels (Boehnke and Phillips 1999; Phillips 1999), thereby resulting in larger and earlier responses. Another possibility could be that the location difference can be detected based on the initial onset slope and thus is earlier than frequency comparison which requires some integration over time if frequencies are in the same critical band.

The difference in ORN latency between △location and $\triangle f_0$ is difficult to reconcile with the notion that ORN indexes perception of concurrent sound objects given that accuracy was comparable in both conditions. However, in the present study, the emphasis was placed on identifying concurrent vowels, which can only occur after an initial segregation of the acoustic data into its constituents. It is possible that △location between the 2 vowels may yield a strong sense of concurrent sound objects even if participants experienced difficulties in identifying the 2 vowels. This would not be reflected in the behavioral data but would be consistent with prior studies showing that the ORN amplitude and latency are related to the perception of concurrent sound objects (Alain 2007).

In addition to the ORN, the effects of $\triangle f_0$ and △location on concurrent speech perception were paralleled by an N2b around 230 ms after sound onset. This N2b is thought to reflect stimulus categorization and decision processes that control behavioral responses in discrimination tasks (Ritter et al. 1979, 1982). In the present study, the N2b may also index a schema-driven process in vowel identification in which the incoming vowels are matched against stored vowel representations (schemata) in working memory.

Listeners' ability to correctly identify both vowels was best accounted for by an additive effect in which the differences of f_0 and spatial location between the 2 vowels are linearly combined together during speech separation and identification. This behavioral advantage observed during the concurrent vowel identification task was paralleled by changes in bilateral auditory cortices that mimic the behavioral effects. Specifically, changes in both accuracy and auditory cortical activity during the ORN and N2b intervals as a function of $\triangle f_0$ or △location were additive such that the sum of $\triangle f_0$ and △location alone equaled the combined effect of having both spectral and spatial differences simultaneously. Previous researchers that have examined the MMN to multidimensional deviant stimuli have revealed additivity during the MMN latency but usually not thereafter (e.g., Schröger 1995; Paavilainen et al. 2001). In the present study, the analyses were carried out on the source waveforms from the N1m generator, which provided only a rough estimate of source activity during the N2b interval.

Using the beamformer approach for neuromagnetic source imaging, we found converging evidence for linear summation of $\triangle f_0$ and △location in auditory cortex. Our results are consistent with those of a functional MRI study which also shows additive effects between pitch and sound source location (Barrett and Hall 2006). It appears that during speech segregation and identification differences in f_0 and spatial location are both accessible and used to enhance the separation and representations of both vowels in sensory memory. The linear combination of f_0 and spatial differences may be used to enhance the perceptual distance between the 2 speech signals, thereby easing the identification of the various sound objects in the mixture. Such a strategy may be particularly beneficial in situations where the acoustic cues promoting segregation are not too salient. In such situations, the listeners may need to integrate all available information to successfully separate co-occurring speech signals. Thus, linear summation may occur in situations of low signal-to-noise ratio where listeners need to accumulate evidence during auditory scene analysis. Further research is needed to assess the extent to which such a strategy depends on the difficulty to separate concurrent speech signals.

The additive effect observed in the present study is interesting in light of current models of information processing and may share some similarity with the redundant signals effect which refers to a behavioral advantage associated with presenting 2 signals that both call for the same response (e.g., Schwarz 1989; Mordkoff and Yantis 1991; Schroter et al. 2007; Miller et al. 2009). In the present context, differences in f_0 and spatial location may constitute 2 different signals that are redundant in the sense that they both involve the same response (i.e., vowel identification). Two general classes of models have been proposed to account for the processing of redundant signals. The horse-race model assumes that separate decision processes are made in parallel for each signal and performance is driven by the quicker process (Mordkoff and Yantis 1991). The coactivation model posits a decision mechanism that takes into account the time needed to process both signals (Mordkoff and Yantis 1993; Miller 2007; Schroter et al. 2007). Our results are more in line with the coactivation model because both $\triangle f_0$ and △location contribute activation to a common decision threshold. They cannot easily be accounted for by a horse-race model in which, presumably, segregation would be based on the most salient cue irrespective of the other cue.

Conclusions

This study shows the contribution of spectral (i.e., f_0) and/or spatial differences to speech separation and how these acoustic differences are combined in primary auditory cortex. Although

spectral and spatial information may be processed into specialized pathways, these acoustic differences are linearly combined to ease the separation and identification of speech sounds. The early processing of spectral and spatial cues may converge in the auditory cortex where their contributions sum together to provide an efficient and optimal processing strategy during the perceptual organization of speech sounds. Further research will help determine the limit of this linear integration and whether it applies to other complex listening situations.

Funding

Canadian Institutes of Health Research and the Natural Sciences and Engineering Research Council of Canada to C.A.; Chinese State Scholarship Fund to Y.D.; "973" National Basic Research Program of China (2009CB320901 to L.L.); National Natural Science Foundation of China (30670704; 30711120563 to L.L.).

Notes

We would like to thank Patricia Van Roon for comments on earlier versions of this manuscript. *Conflict of Interest*: None declared.

References

Ahmed B, Garcia-Lazaro JA, Schnupp JW. 2006. Response linearity in primary auditory cortex of the ferret. J Physiol. 572:763–773.

Alain C. 2007. Breaking the wave: effects of attention and learning on concurrent sound perception. Hear Res. 229:225–236.

Alain C, McDonald KL. 2007. Age-related differences in neuromagnetic brain activity underlying concurrent sound perception. J Neurosci. 27:1308–1314.

Alain C, Reinke K, He Y, Wang C, Lobaugh N. 2005. Hearing two things at once: neurophysiological indices of speech segregation and identification. J Cogn Neurosci. 17:811–818.

Assmann PF, Summerfield Q. 1990. Modeling the perception of concurrent vowels: vowels with different fundamental frequencies. J Acoust Soc Am. 88:680–697.

Assmann PF, Summerfield Q. 1994. The contribution of waveform interactions to the perception of concurrent vowels. J Acoust Soc Am. 95:471–484.

Barrett DJ, Hall DA. 2006. Response preferences for "what" and "where" in human non-primary auditory cortex. Neuroimage. 32:968–977.

Bizley JK, Walker KM, Silverman BW, King AJ, Schnupp JW. 2009. Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. J Neurosci. 29:2064–2075.

Boehnke SE, Phillips DP. 1999. Azimuthal tuning of human perceptual channels for sound location. J Acoust Soc Am. 106:1948–1955.

Chalikia MH, Bregman AS. 1989. The perceptual segregation of simultaneous auditory signals: pulse train segregation and vowel segregation. Percept Psychophys. 46:487–496.

Chau W, McIntosh AR, Robinson SE, Schulz M, Pantev C. 2004. Improving permutation test power for group analysis of spatially filtered MEG data. Neuroimage. 23:983–996.

Cheyne D, Bakhtazad L, Gaetz W. 2006. Spatiotemporal mapping of cortical activity accompanying voluntary movements using an event-related beamforming approach. Hum Brain Mapp. 27:213–229.

Cox RW. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res. 29:162–173.

Drennan WR, Gatehouse S, Lever C. 2003. Perceptual segregation of competing speech sounds: the role of spatial location. J Acoust Soc Am. 114:2178–2189.

Gondan M, Roder B. 2006. A new method for detecting interactions between the senses in event-related potentials. Brain Res. 1073–1074:389–397.

Holmes CJ, Hoge R, Collins L, Woods R, Toga AW, Evans AC. 1998. Enhancement of MR images using registration for signal averaging. J Comput Assist Tomogr. 22:324–333.

Levanen S, Hari R, McEvoy L, Sams M. 1993. Responses of the human auditory cortex to changes in one versus two stimulus features. Exp Brain Res. 97:177–183.

Machens CK, Wehr MS, Zador AM. 2004. Linearity of cortical receptive fields measured with natural sounds. J Neurosci. 24:1089–1100.

Miller J. 2007. Contralateral and ipsilateral motor activation in visual simple reaction time: a test of the hemispheric coactivation model. Exp Brain Res. 176:539–558.

Miller J, Beutinger D, Ulrich R. 2009. Visuospatial attention and redundancy gain. Psychol Res. 73:254–262.

Mordkoff JT, Yantis S. 1991. An interactive race model of divided attention. J Exp Psychol Hum Percept Perform. 17:520–538.

Mordkoff JT, Yantis S. 1993. Dividing attention between color and shape: evidence of coactivation. Percept Psychophys. 53:357–366.

Paavilainen P, Valppu S, Naatanen R. 2001. The additivity of the auditory feature analysis in the human brain as indexed by the mismatch negativity: 1+1 approximately 2 but 1+1+1 <3. Neurosci Lett. 301:179–182.

Phillips DP. 1999. Auditory gap detection, perceptual channels, and temporal resolution in speech perception. J Am Acad Audiol. 10:343–354.

Remez RE, Ferro DF, Wissig SC, Landau CA. 2008. Asynchrony tolerance in the perceptual organization of speech. Psychon Bull Rev. 15:861–865.

Remez RE, Rubin PE, Berns SM, Pardo JS, Lang JM. 1994. On the perceptual organization of speech. Psychol Rev. 101:129–156.

Ritter W, Simson R, Vaughan HG Jr, Friedman D. 1979. A brain event related to the making of a sensory discrimination. Science. 203:1358–1361.

Ritter W, Simson R, Vaughan HG Jr, Macht M. 1982. Manipulation of event-related potential manifestations of information processing stages. Science. 218:909–911.

Robinson SE. 2004. Localization of event-related activity by SAM(erf). Neurol Clin Neurophysiol. 2004:109.

Robinson SE, Rose DF. 1992. Current source estimation by spatially filtered MEG. In: Romani G, editor. Biomagnetism: clinical aspects. Amsterdam (The Netherlands): Excerpta Medica. p. 761–765.

Robinson SE, Vrba J. 1998. Functional neuroimaging by synthetic aperture magnetometry (SAM). In: Yoshimoto T, Kotani M, Kuriki S, Karibe H, Nakasato N, editors. Recent advances in biomagnetism. Sendai (Japan): Tohuku University Press. p. 302–305.

Rossi-Katz J, Arehart KH. 2009. Message and talker identification in older adults: effects of task, distinctiveness of the talkers' voices, and meaningfulness of the competing message. J Speech Lang Hear Res. 52:435–453.

Schröger E. 1995. Processing of auditory deviants with changes in one versus two stimulus dimensions. Psychophysiology. 32:55–65.

Schröger E. 1998. Measurement and interpretation of the mismatch negativity. Behav Res Methods Instrum Comput. 30:131–145.

Schroter H, Ulich R, Miller J. 2007. Effects of redundant auditory stimuli on reaction time. Psychon Bull Rev. 14:39–44.

Schwarz W. 1989. A new model to explain the redundant-signals effect. Percept Psychophys. 46:498–500.

Shackleton TM, Meddis R. 1992. The role of interaural time difference and fundamental frequency difference in the identification of concurrent vowel pairs. J Acoust Soc Am. 91:3579–3581.

Spieth W, Curtis JF, Webster JC. 1954. Responding to one of two simultaneous messages. J Acoust Soc Am. 26:391–396.

Summerfield Q, Assmann PF. 1991. Perception of concurrent vowels: effects of harmonic misalignment and pitch-period asynchrony. J Acoust Soc Am. 89:1364–1377.

Takegata R, Morotomi T. 1999. Integrated neural representation of sound and temporal features in human auditory sensory memory: an event-related potential study. Neurosci Lett. 274:207–210.

Treisman AM. 1964. Verbal cues, language, and meaning in selective attention. Am J Psychol. 77:206–219.

Van Veen BD, van Drongelen W, Yuchtman M, Suzuki A. 1997. Localization of brain electrical activity via linearly constrained

minimum variance spatial filtering. IEEE Trans Biomed Eng. 44:867–880.
Wenzel EM, Arruda M, Kistler DJ, Wightman FL. 1993. Localization using nonindividualized head-related transfer functions. J Acoust Soc Am. 94:111–123.
Wightman FL, Kistler DJ. 1989a. Headphone simulation of free-field listening. I: Stimulus synthesis. J Acoust Soc Am. 85:858–867.
Wightman FL, Kistler DJ. 1989b. Headphone simulation of free-field listening. II: Psychophysical validation. J Acoust Soc Am. 85: 868–878.
Wolff C, Schröger E. 2001. Human pre-attentive auditory change-detection with single, double, and triple deviations as revealed by mismatch negativity additivity. Neurosci Lett. 311:37–40.
Woods DL, Alain C. 1993. Feature processing during high-rate auditory selective attention. Percept Psychophys. 53:391–402.
Woods DL, Alain C. 2001. Conjoining three auditory features: an event-related brain potential study. J Cogn Neurosci. 13:492–509.
Woods DL, Alho K, Algazi A. 1994. Stages of auditory feature conjunction: an event-related brain potential study. J Exp Psychol Hum Percept Perform. 20:81–94.