Salminen, Nelli H.; Takanen, Marko; Santala, Olli; Alku, Paavo; Pulkki, Ville

Neural realignment of spatially separated sound components

Published in:
Journal of the Acoustical Society of America

DOI:
10.1121/1.4921605

Published: 01/01/2015

Please cite the original version:
I. INTRODUCTION

In our conscious experience, everyday auditory environments are organized into separate auditory objects, such as a clock ticking on the wall, a dog barking outside, and a colleague talking in the corridor. However, when these sounds arrive to the ears, they overlap in the same waveform. How does the brain decide which components in this mixture belong to the same object and which do not? The auditory system performs this task of auditory scene analysis using various physical cues. For instance, sounds with the same fundamental frequency and a concurrent onset are often deemed to belong to the same object (Hukin and Darwin, 1995; Darwin and Hulkin, 1998, 1999; Griffins and Warren, 2004; Shackleton and Meddis, 1992; Bizley and Cohen, 2013). In natural environments, sound components from the same source also tend to arrive from the same location. Yet, spatial location has relatively little influence on the grouping of concurrent sounds (Shackleton and Meddis, 1992; Shackleton et al., 1994; Culling and Summerfield, 1995; Darwin, 2008; Schwartz et al., 2012). A spatial separation between two sound sources does not prevent them from being perceptually fused into a single object when other cues are in favor of grouping (Hukin and Darwin, 1995; Darwin and Hulkin, 1998, 1999; Shinn-Cunningham et al., 2007; Takanen et al., 2013). This fusion makes the identification and localization of concurrently presented sounds a challenging task (Woods and Colburn, 1992; Best et al., 2005; Best et al., 2007; Lee et al., 2009; Croghan and Grantham, 2010; Schwartz et al., 2012). Here, we conducted a magnetoencephalography (MEG) study to better understand the brain mechanisms underlying the difficulties in utilizing spatial location as a segregation cue for concurrent sounds.

The relative weights given to various grouping cues have been systematically studied with ambiguous stimuli that can be interpreted as either one or two sound objects depending on which cues the listener primarily relies on. For instance, if speech is divided into a low- and a high-frequency band and these two are presented from different locations, only one stream of speech is heard (Broadbent and Ladefoged, 1957; Darwin and Hulkin, 2004). The fundamental frequency, F0, provides a more salient cue for separating competing sound sources than the spatial location (Shackleton and Meddis, 1992)—concurrent utterances of speech differing in F0 are not fused even when they are colocated (Darwin and Hulkin, 2004). However, the difference in F0 must be large enough (Cutting, 1976; Culling and Darwin, 1993; Bird and Darwin, 1998). The relative weights of different grouping cues can also be evaluated with simpler stimuli, for instance, with tone complexes from which a single tonal component is segregated by mistuning, asynchronous onset, or location (McDonald and Alain, 2005; Lipp et al., 2010; Kocsis et al., 2014). In such situations, both mistuning and asynchronous onset lead to the perception of two separate objects while a spatial separation is less likely to lead to segregation. The segregation of a sound into two objects also has a neural correlate in the event-related potentials. The perception of not only one but of two objects is associated with a negative response, the object-related negativity (Alain et al., 2001;
Alain et al., 2002; McDonald and Alain, 2005; Sanders et al., 2008; Sanders et al., 2011; Lipp et al., 2010; Du et al., 2011; Weise et al., 2012; Kocsis et al., 2014), and the amplitude of this response is correlated with the likelihood of successful sound source segregation (McDonald and Alain, 2005; Sanders et al., 2008; Sanders et al., 2011; Du et al., 2011).

Auditory grouping can also be evaluated with stimuli whose identity depends on whether its components are grouped across spatial locations or not (Hukin and Darwin, 1995; Darwin and Hulkin, 1998, 1999; Shinn-Cunningham et al., 2007). For this purpose, a single sound stimulus is divided into two components and these two are presented from different locations. Fusion across locations can then be inferred from identification: if both sound components contribute to the perceived identity of the resulting object, fusion has taken place. For instance, if a single tone is spatially separated from a tone replica of a vowel sound, the tone continues to contribute to the perceived identity of the vowel despite a spatial separation (Hukin and Darwin, 1995; Darwin and Hulkin, 1998, 1999; Shinn-Cunningham et al., 2007). In contrast, if a single tone is mistuned or it has a different onset time than the rest of the complex, it contributes less to the perceived identity. As another example, a vowel sound can be divided into two components: one containing the odd and the other containing the even-formant components (Takanen et al., 2013), the resonances that determine vowel identity. Neither component alone is sufficient for vowel identification. Yet, when these two components are presented together, the vowel is identified with ease independent of the extent of spatial separation between the components. Together, these findings show that concurrent sound sources are not perceptually separated from one another based on spatial location when other cues favor fusion.

Here, rather than trying to establish the conditions under which perceptual fusion does or does not take place, the aim was to understand how the conflicting location information is processed in the brain when fusion across locations does occur. When considering the cortical representation of auditory space, the apparent ease of fusing spatially separated sounds seems puzzling. In the auditory cortex, neural tuning to spatial location is a ubiquitous property: in some cortical fields, spatially selective neurons constitute >50% of all neurons (Tian et al., 2001; Woods et al., 2006). Auditory cortical neurons are widely and laterally tuned with their receptive fields spanning one hemifield of the auditory space (Woods et al., 2006; Werner-Reiss and Groh, 2008; Salminen et al., 2009; Salminen et al., 2012). Two concurrently presented sounds, one from the left and the other from the right side of the midline, would then activate separate neural populations, and thereby their segregation should be possible based on the spatial separation. Yet, psychoacoustical studies show that this segregation is not straightforward (Shackleton and Meddis, 1992; Woods and Colburn, 1992; Shackleton et al., 1994; Culling and Summerfield, 1995; Best et al., 2005, 2007; Darwin, 2008; Lee et al., 2009; Schwartz et al., 2012). A potential explanation for this is that the identity and location information are processed separately in the auditory cortex as suggested by the dual stream hypothesis (Rauschecker and Tian, 2000; Arnott et al., 2004; Ahveninen et al., 2006; Ahveninen et al., 2013). According to this view, the identification of auditory objects relies on the ventral “what” stream and localization on the dorsal “where” stream. Then, the location information might not be available in the object construction and identification process presumably taking place in the ventral stream.

Alternatively, the auditory nervous system could solve the spatial conflict by realigning the sound components depending on which auditory object they belong to. That is, the auditory cortex would represent all sound components belonging to the same auditory object within the same population of spatially selective neurons, regardless of their physical locations. For instance, the spatially separated odd- and even-formant components of a vowel sound (Takanen et al., 2013) would be treated in the brain so that the representations of the individual components would be shifted to match the perceived location of the fused vowel object. Then, from the point of view of the auditory cortex, there would be no more spatial conflict and the identification and localization of the resulting vowel would be straightforward. Such a realignment process would make the original spatial information unavailable at the level of the auditory cortex. If such a process takes place in the brain, it would explain not only how spatially separated objects can be fused back together, but also why the segregation of concurrent sounds is not significantly facilitated by spatial separation.

The purpose of this study was to test whether the auditory cortex performs the above-hypothesized spatial realignment of sound components. To this end, we recorded auditory cortical responses in MEG to a stimulus where the odd- and even-formant components of a vowel sound are simultaneously presented from opposite sides of the midline (Fig. 1). These spatially separated vowels are perceived as a vowel sound originating from a single location, showing that the components are perceptually placed into the same spatial location (Takanen et al., 2013). The present experiment aimed to reveal how this perceptual reorganization is reflected in the cortical representation of the sound components. Specifically, it was evaluated which spatially selective population, the left- or right-tuned one, each component activates in the cortex. To this end, the experimental design was based on a stimulus-specific adaptation paradigm that

FIG. 1. The aperiodic vowel stimulus. (A) The Finnish vowel /æ/ was divided into two components containing the odd and even formants. (B) These two components were presented from different locations. (C) This stimulus leads to the percept of an identifiable vowel originating from the location of the odd-formant component.
capitalizes on the sensitivity of auditory cortical responses to stimulation history (Butler, 1968, 1972; McEvoy et al., 1993; Alveninen et al., 2006; Salminen et al., 2009; Salminen et al., 2010). For instance, cortical N1 responses to a probe sound at the right are smaller in amplitude when the probe is preceded by an adaptor at the right than when the adaptor is at the left (Butler, 1972; McEvoy et al., 1993; Salminen et al., 2009). This effect is also frequency specific: adaptors with frequencies near to the probe cause stronger attenuation than adaptors with far frequencies (Butler, 1968; Näätänen et al., 1988). Here, the spatial representations of the odd- and even-formant components of the vowel stimulus were investigated by using the isolated odd- and even-formant components as adaptor sounds and the entire spatially separated vowel as the probe. Specifically, the present experiment tested whether the spatial representations of the individual components correspond to their objective, physical locations or to the subjective, perceived location of the vowel object resulting from the fusion. If the latter case shows to be true, this would suggest that the human brain performs spatial realignment of concurrent sound components.

II. METHODS

A. Stimulus preparation

The stimulus was an aperiodic vowel-like sound divided into two components containing the odd or the even formants [Fig. 1(A)]. These components were presented from different directions [Fig. 1(B)]. Additionally, a neutral stimulus without a clear phonemic identity was generated, and similarly divided into two components. The components in both the vowel and the neutral stimulus had a common onset that is a strong grouping cue (Griffins and Warren, 2004; Bizley and Cohen, 2013). Head-related transfer functions (HRTFs) were used for spatial sound presentation with headphones, a necessity in the MEG recordings.

First, Finnish vowels (/æ/, /ɛ/, /i/, /o/, /ɑ/., /ɒ/, /æ/, and /ɛ/) produced by a male speaker were recorded in anechoic conditions using a condenser microphone (B&K 4188, Brüel & Kjaer, Naerum, Denmark). The speech signals were sampled at 22.05 kHz, using a resolution of 16 bits, and the sampling rate was later reduced to 16 kHz in order to efficiently apply linear prediction (LP) based algorithms to generate the stimuli (Rabiner and Schafer, 1978). Each of the recorded vowels was processed with an automatic inverse filtering algorithm (Alku et al., 1999) in order to separate each vowel into two parts: (1) the derivative of the glottal flow waveform, denoted as a discrete-time signal by \( g(n) \), and (2) the vocal tract transfer function, denoted in z-domain by \( V_{\text{nat}}(z) \). For modeling the production of the neutral stimulus, an artificial vocal tract transfer function, \( V_{\text{neut}}(z) \), was constructed. This was accomplished by modeling the acoustics of an ideal lossless tube with a uniform cross-section having the average male vocal tract length (17 cm) and using 350 m/s as the speed of sound (Rabiner and Schafer, 1978). This procedure yielded \( V_{\text{neut}}(z) \), whose four lowest formants had the center frequencies of 515 Hz, 1545 Hz, 2575 Hz, and 3605 Hz having a constant difference of 1030 Hz. The bandwidth of each formant in the neutral stimulus was set to 218 Hz, the value which corresponded to the average formant bandwidth computed from \( V_{\text{nat}}(z) \) of the vowel /æ/. The resonances of the vocal tract transfer functions, \( V_{\text{nat}}(z) \) and \( V_{\text{neut}}(z) \), were enhanced using 0.4 as the value of the enhancement parameter (Ling et al., 2006). This step aimed to improve vowel identification by emphasizing the formant amplitudes and by attenuating the spectral valleys between the formants while keeping the center frequencies of the formants unchanged. The vocal tract transfer functions following the formant enhancement are denoted below as \( \hat{V}_{\text{nat}}(z) \) and \( \hat{V}_{\text{neut}}(z) \).

To acquire stimuli with a denser spectral structure than that of phonated vowels, aperiodic noise sequences were used as excitation to the vocal tract transfer functions. However, the spectral shape of the noise sequences were matched to that of \( g(n) \). Specifically, a second-order LP analysis was performed on the original \( g(n) \) estimated from the vowel /æ/, and a set of incoherent white-noise signals were filtered with the same LP filter to obtain aperiodic noise sequences with the desired spectral envelope. Moreover, 18 incoherent white-noise signals were filtered with the same LP filter. The resulting noise sequences were later used to generate the even- and odd-formant components of the eight vowels, and the corresponding components of the neutral stimulus. Finally, noise-excited counterparts for the original vowels and the neutral stimulus were obtained by filtering the different noise sequences either with \( \hat{V}_{\text{nat}}(z) \) of the corresponding vowels or \( \hat{V}_{\text{neut}}(z) \). The effect of lip-radiation was also emulated by processing the signals with a high-pass filter (Fant, 1970).

Subsequently, the noise-excited signals were filtered with a set of bandpass filters to separate them into odd- and even-formant components. Specifically, a 180th-order window-based finite impulse response filter was generated around each formant based on the magnitude response of the corresponding vocal tract transfer function, and the signals were processed with such filters to obtain the odd- and even-formant components. Two incoherent noise-sequences were used to generate the components, separately for each signal, so that they could not be grouped together based on the coherence. The 1/16 octave band smoothed magnitude responses of the two components of the vowel /æ/ are illustrated in Fig. 1(A).

Finally, the signals were processed with HRTFs to simulate sound source locations in headphone reproduction. Prior to this, the signals were resampled at 48 kHz, which corresponded to the sampling rate of the HRTF database. Measured HRTFs of a dummy head (Cortex MK2, Cortex Instruments, Regensburg, Germany) were used. To obtain a symmetric HRTF database, the right-ear transfer function for a given angle was replaced with the left-ear transfer function corresponding to the same angle on the opposite side of the midline. The sound source locations used here were \(-60^\circ, 0^\circ, \) or \(60^\circ\) on the horizontal plane.

B. Psychoacoustical validation

A psychoacoustical experiment was performed to confirm that similar results are obtained using headphone
reproduction as in a previous experiment conducted in free-field conditions (Takanen et al., 2013). Specifically, the aim was to verify that the vowel /æ/ is correctly identified despite the spatial separation between the components, and that neither the odd nor the even component alone is sufficient for identification. Furthermore, the perceived location of the vowel was evaluated. Finally, it was tested whether the neutral stimulus is identified as any specific vowel.

The experiment was conducted in a single-walled sound-isolated listening booth using circumaural headphones (Sennheiser HD 650, Wedemark, Germany) and a mouse-operated graphical user interface. Thirteen native Finnish speakers (age mean = 28, standard deviation = 5, three females) were recruited from the students and staff of Aalto University. The subjects were not aware of the purpose of the experiment. Before starting the experiment, the participants were given the opportunity to listen to the eight Finnish vowel stimuli as many times as they deemed necessary. After listening to each vowel at least once, they were allowed to begin the test.

A total of 15 stimuli were included in the listening experiment. Nine of these were intact stimuli (the eight Finnish vowels and the neutral) presented from a single location (randomly chosen from −60°, 0°, and 60°). Two stimuli were spatially separated: the odd- and even-formant components of /æ/ or the neutral stimulus presented concurrently from different locations (−60° and 60° randomly assigned to the components). Finally, four stimuli were odd or even components of /æ/ or the neutral stimulus presented in isolation from 0°. Stimulus duration was 500 ms. The experiment consisted of 75 trials, resulting from the 15 test cases being measured 5 times. In each trial, the task of the participant was twofold. First, they heard the stimulus once and were asked to identify it as one of the eight Finnish vowels. Alternatively, they could answer that the stimulus was not a vowel. Second, and only if the stimulus was identified as a vowel, they heard the same stimulus again and indicated whether they heard the vowel from the left, center, or right. The presentation order of the trials was randomized for each participant. On average, the participants completed the experiment in 8 min.

The spatially separated vowel, /æ/, was correctly identified with >85% accuracy, but the identification rate with the isolated odd and even formants was at the chance level of 11.1% \( \chi^2 = 99.5, \text{df} = 2, p < 0.001 \). For evaluating the perceived location of the spatially separated vowel, the localization responses of the participants were coded with values of −1, 0, and 1 for the direction of the even component, the center, and for the direction of the odd component, respectively. The average perceived location in this scale was greater than zero \( y = 0.3, \sigma = 0.53, t(12) = 4.19, p < 0.001 \), which verified that the perceived direction of the vowel was to the direction of the odd component. The neutral stimulus was most commonly perceived as /æ/, /u/, or “not a vowel.” Therefore, a t-test was employed to analyze whether the neutral stimulus was more likely to be identified as either /æ/ or /u/. The result was nonsignificant, indicating that the neutral stimulus was not systematically identified as any particular vowel.

C. MEG

Twelve native Finnish speakers (age mean = 27, standard deviation = 5, three females) took part in the MEG experiment with written informed consent. All experimental procedures were approved by the ethical committee of Aalto University.

A stimulus-specific adaptation paradigm (Butler, 1972; Salminen et al., 2009) was used in the MEG recordings to test how the auditory cortex represents the spatially separated vowel. In this paradigm, a probe sound is presented in an alternating sequence with various adaptor sounds. Crucially, responses are recorded exclusively to the probe and the probe is kept constant. Then, all variations in response amplitude can be attributed to the influence of the adaptor. When the adaptor is presented from the same location as the probe, response amplitudes are smaller than when the two sounds are presented from different locations. This is presumably due to the spatial selectivity of the single neurons underlying the brain response. The experimental design here relies also on the frequency specificity of the N1 adaptation (Butler, 1968; Näätänen et al., 1988), that is, on stronger adaptation occurring between adaptor and probe components with matching rather than with non-matching frequencies.

The stimulus-specific adaptation paradigm was used in order to test which set of spatially selective neurons each component of the spatially separated stimulus activates, the left- or the right-tuned. The spatially separated vowel, /æ/, was used as the probe and the isolated odd- and even-formant components as adaptors. The probe had the odd-formant component in the right (60°) and even-formant component in the left (−60°) side, or vice versa. The adaptors were presented from either the left or the right side, i.e.,
Even component in the probe [Figs. 2(A) and 3(A)]. The adaptor location matched either that of the odd or the even component in the probe (gray). This is in agreement with the prediction based on the even-component being spatially realigned with the odd component and thereby to the perceived vowel location.

FIG. 3. The neural representation of the even-formant component matched the perceived rather than the physical location of the component. (A) The spatially separated vowel served again as the probe and the adaptor was the isolated even-formant component. (B) If the neural representation corresponds to the physical sound scene, the strongest adaptation should result from the even component being at the same location in the adaptor and the probe (top). The two adaptor locations could result in equally strong adaptation if the even-formant component is represented partially in the two locations: the actual direction of the component and the perceived direction of the odd component (middle). Alternatively, the adaptor location matching with the perceived direction of the even component could lead to the strongest adaptation, provided that the even-formant component is realigned with the odd-formant one in the neural representation (bottom). (C) The left-hemispheric N1 response was attenuated more when the adaptor location corresponded to the physical sound scene, the strongest adaptation should result from the even component being at the same location in the adaptor and the probe (gray). This is in agreement with the prediction based on the even-formant component being spatially realigned with the odd component and thereby to the perceived vowel location.

The adaptor location matched either that of the odd or the even component in the probe [Figs. 2(A) and 3(A)].

There were altogether four adaptor conditions: the odd or even formants presented from the left or the right location. The probe and adaptor sounds were presented in an alternating sequence with an (onset to onset) interstimulus-interval (ISI) of 1 s. Stimulus duration was 200 ms. In an additional reference condition, no adaptors were presented, i.e., the probe was presented with an ISI of 2 s. Each adaptor condition was presented in a separate stimulation block resulting in, altogether, five blocks for the vowel and the neutral stimulus. Each stimulus block was presented for about 5 min while the subject focused on reading a self-selected text under the instruction to sit still and ignore the sound stimulation. The sounds were presented with a MEG-compatible tube-phone system with insert tips. This system is capable of preserving the stimulus characteristics within the frequency range from 300 to 7000 Hz. For six subjects, the probe had the odd component in the right and the even component in the left location [Figs. 2(A) and 3(A)]. For the remaining six subjects, the directions were reversed [Fig. 4(A)]. This was done to control for potential differences between cortical hemispheres.

MEG was recorded with a whole-head device (Vectorview, Elekta-Neuromag, Helsinki, Finland) with 102 sensor elements, each consisting of two planar gradiometers and a magnetometer. The recording sample rate was 1024 Hz and the bandpass 0.01–300 Hz. Before the recordings, five head position indicator coils were attached to the scalp and the positions of the coils, three anatomical landmarks, and ~20 additional points along the scalp were digitized. The position of the head was estimated before the onset of each stimulation block for later use in the data analysis. Blinks and eye-movements were monitored with vertical and horizontal electro-oculogram. The data were filtered at 1–30 Hz and event-related fields (ERFs) were obtained offline for a period from 100 ms before to 500 ms after probe onset. Epochs containing deviations >3000 fT/cm in gradiometers or 150 μV in the electro-oculogram were discarded. This resulted in a minimum of 100 repetitions averaged for each stimulus condition.

Further analysis focused on the amplitude of the N1, a response peak arising in MEG recordings from the auditory cortex at ~100 ms after sound onset (Hari et al., 1980; Hämäläinen et al., 1993). The neural activity at the level of the auditory cortical source was modeled with equivalent current dipoles (ECDs; Hämäläinen et al., 1993) with the Xfit software (Neuromag, Helsinki, Finland). A 100-ms pre-stimulus baseline and a spherical head model were used. The analysis was based on a set of 22 gradiometer pairs above the temporal lobe for the left and the right hemispheres separately. First, a model for the N1 source was obtained based on the reference condition in which no adaptors were presented.

FIG. 4. The results generalized across hemispheres. (A) To ensure that the effect was not specific to the spatial arrangement and to the left hemisphere, the recordings were repeated for a stimulus in which the odd- and even-formant component locations were reversed. (B) For both odd- (top) and even-formant (bottom) adaptors, the N1 response amplitude was more attenuated by adaptors in the perceived direction of the vowel sound (black) than by adaptors on the opposite side (gray). The location-specific adaptation occurred in the right hemisphere that was for this stimulus contralateral to the perceived vowel location.
An ECD was fit at 1 ms intervals to the ERF and the N1 was identified as a magnitude peak within 80–150 ms from stimulus onset. The location and orientation of the source at this time point was then used as the model for the N1 source. With this procedure, a reliable model could be obtained for every subject and for both cortical hemispheres (goodness of fit on the average 91%). These coordinates were then used for modeling the activity in other conditions by keeping the location and orientation fixed while letting the source strength vary. The N1 was identified as a peak in the resulting source strength within 80–150 ms latency. The ECD analysis was performed for the vowel and the neutral stimulus, and for each subject separately. For illustration (Fig. 5), the N1 response amplitudes were normalized for each subject with respect to the average and standard deviation across adaptor conditions. As the analysis aimed at quantifying the amplitude of the N1 response rather than the source location or orientation, the analysis could have alternatively been performed in the sensor space. To ensure that the results were not sensitive to the choice of analysis procedure, the N1 amplitudes were also quantified from the root-mean-square of the 22 gradiometer pairs. This analysis provided results highly similar to those obtained from the ECD and are therefore not reported.

A repeated-measures analysis of variance with repeating factors adaptor type (odd vs even), adaptor location (same as vs different from the odd component in the probe), and hemisphere (contralateral vs ipsilateral to the odd component in the probe) was performed on the N1 peak amplitudes to the probe sound. The data of the two subject groups with opposite sound source locations were combined. There was no evidence for differences between the two groups [repeated-measures analysis of variance with subject group as a categorical predictor $F(1,10) < 2.67, p > 0.1$ for all main effects and interactions].

**III. RESULTS**

Auditory cortical N1 responses were recorded to a vowel sound divided into two components: odd-formant component presented from the right and even-formant component from the left side of the midline (Fig. 1). These vowels are identified with ease and localized consistently to the direction of the odd-formant component (Takanen et al., 2013; see also Sec. II). The spatially separated vowel was used as a probe stimulus and was preceded by an adaptor consisting of either the odd- or the even-formant component presented in isolation. Adaptors at the same location are known to cause more attenuation than adaptors at a different location (Butler, 1972; Salminen et al., 2009). Then, for instance, the odd-formant adaptor would be expected to cause stronger adaptation when its location matches the location of the odd-formant component in the probe than when the locations differ [Fig. 2(B)]. The amplitude of the N1 response depended on the location of the adaptor sound [hemisphere $\times$ adaptor type $\times$ location interaction $F(1,11) = 10.6; p = 0.0076$]. For the odd-formant adaptor, the left-hemispheric brain responses followed the above prediction: the adaptor on the right side caused stronger adaptation than that on the left [Fig. 2(C)]. In other words, adaptation was stronger when the odd-formant components in the adaptor and probe were at the same location than when their locations differed (post hoc comparison for odd-formant adaptor location, $p = 0.0002$).

For the even-formant adaptor, the expected outcome depends on whether the spatial representation of the even-formant component corresponds to the physical source location on the left or to the perceived location of the vowel object on the right. If the neural representation corresponds to the actual location, the strongest adaptation should follow from the adaptor at the left location, i.e., the actual location of the even-formant component in the probe [Fig. 3(B)]. Alternatively, if the auditory nervous system performs spatial realignment and positions the even-formant component in the probe with the perceived location of the vowel in its internal representation, the strongest adaptation should result from the adaptor being on the right. Finally, the change in the representation could be partial. Then, the result would be intermediate with the two adaptor locations causing similar levels of adaptation. The effect of the even-formant adaptor followed the prediction of a complete shift in the spatial representation to the perceived location [Fig. 3(C)]. The strongest adaptation was caused by the adaptor presented at the perceived object location (post hoc comparison for even-formant adaptor location, $p = 0.0015$). This signifies that at the stage of the auditory cortex, the neural representation of the even-formant component in the probe had shifted from the physical location of the sound component to the perceived location of the auditory object.

In the above, the N1 amplitude was affected by adaptor location only in the left hemisphere, that is, the hemisphere contralateral to the perceived location of the vowel object. This is consistent with previous studies showing larger location-dependent adaptation effects in the hemisphere contralateral to the stimulus location (Salminen et al., 2009; Salminen et al., 2010). However, to ensure that the effect...
observed here is not specific to the hemisphere, the sound source locations were inverted for a subset of the participants: the odd-formant component was presented on the left and the even-formant on the right side [Fig. 4(A)]. The effect generalized across cortical hemispheres [spatial arrangement as a categorical predictor $F(1,10) < 2.67; p > 0.1$ for all main effects and interactions; Fig. 5(A)]. Strongest adaptation occurred again for the adaptors at the perceived location of the vowel object, this time at the left [Fig. 4(B)]. This occurred in the right hemisphere contralateral to the perceived object location.

The above results were obtained with a speech-like sound whose phonemic identity could be identified by the subjects with ease. In other words, the fusion of the two sound components resulted in a familiar auditory object. To evaluate the importance of phonemic identity in the fusion process, the experiment was repeated for a sound with a formant structure that did not correspond to any particular vowel. This sound was also divided into its odd- and even-formant components that were, thereafter, presented from different locations. The brain responses to this stimulus behaved similarly to those obtained with the identifiable vowel sound [Fig. 5(B)]. The adaptors at the location of the odd-formant component caused stronger adaptation than those in the even-formant component location [hemisphere $\times$ adaptor location interaction $F(1,10) = 10.9; p = 0.0071$; post hoc comparison in contralateral hemisphere for location, $p = 0.0007$]. This applied to both the odd- and the even-formant adaptors. Further, the location-dependent adaptation occurred in the hemisphere contralateral to the odd-formant component in the probe. Therefore, the spatial realignment of the two components occurred also in the absence of clear object identity.

Finally, to assess potential differences between the odd- and even-formant adaptors in the magnitude of adaptation, the response amplitudes were pooled across hemisphere and adaptor location and the resulting N1 amplitudes were compared. This analysis was performed in order to evaluate whether the results might simply be explained through greater adaptation occurring with the odd-formant adaptor. The overall magnitude of adaptation did not differ between the odd- and even-formant adaptors [$F(1,11) = 0.01, p = 0.92$]. Also, the N1 response amplitudes to the odd- and even-formant adaptors themselves were of similar magnitudes [$F(1,11) = 1.92, p = 0.19$].

IV. DISCUSSION

Here, a MEG study was conducted in order to understand how the brain processes location information when sound components are distributed across different locations. To this end, brain responses were recorded to a spatially separated vowel stimulus that is perceptually fused despite the spatial conflict. The results showed that the cortical representation of the sound did not reflect the original conflicting locations of the individual sound components, but rather the perceived location of the vowel object resulting from the fusion. The present results suggest that the auditory nervous system resolves the spatial conflict by the stage of the auditory cortex and within the first 100 ms after sound onset. These findings support the hypothesis that the human brain implements a spatial realignment mechanism that places sound components to the same spatial location if they are deemed to belong to the same object. Apart from explaining how spatially separated sounds can be fused back together, the neural mechanism of spatial realignment may also contribute to the difficulties experienced by human listeners in identifying and localizing concurrently presented sounds. The spatial realignment proposed here would eliminate the spatial separation cues that could potentially facilitate the segregation of the concurrent sources. This lack in segregation cues would then result in difficulties in identification since the components belonging to each source are intermingled in the same object. Also, since location information is manipulated in the grouping process, this may give rise to biased localization of the individual sound components.

The N1 responses recorded here reflect neural activity pooled across auditory cortical areas. Therefore, based on the present data, it cannot be specified which auditory fields represented the perceived rather than physical locations for the vowel stimulus. It can, however, be concluded that the conflict between spatial and other grouping cues was resolved no later than in the auditory cortex and within the first 100 ms after sound onset. This has implications to the stage at which spatial information is linked to auditory objects, that is, to the stage at which spatial locations are assigned to auditory objects. For the spatially separated vowel stimulus, identification could, in principle, be achieved by simply ignoring all spatial information and by handling localization at a later processing stage. Such a processing strategy would be consistent with the dual stream model of auditory cortical processing (Rauschecker and Tian, 2000; Arnott et al., 2004; Ahveninen et al., 2006; Ahveninen et al., 2013): the ventral stream would construct and identify the vowel and this object would be linked to its spatial location at a later stage. The spatial conflict arises and thereby the spatial realignment becomes necessary only when location is associated to the fused object. Therefore, the spatial realignment can be interpreted as a marker for location assignment. Since the alignment was established by the 100 ms latency of the auditory cortical N1 response, our findings suggest that locations are assigned to auditory objects at a relatively early stage in the auditory cortex or perhaps even subcortically. This signifies that the representation of sound source location is closely linked to objects in the auditory cortex. This is consistent with single-unit recordings showing that spatial sensitivity is a ubiquitous property in the auditory cortex: selectivity to location is found in all auditory cortical fields and it overlaps with sensitivity to other sound features at the level of single neurons (Tian et al., 2001; Middlebrooks et al., 2002; Woods et al., 2006; Bizley et al., 2009). Recent work suggests that horizontal sound source location is represented in the human auditory cortex as a population rate code consisting of two opponent populations (Salminen et al., 2009; Salminen et al., 2010; Magezi and Krumbholz, 2010; reviewed in Salminen et al., 2012; Briley et al., 2013). Such coding of location is well suited for overlapping representations of multiple sound features (Stecker et al., 2005).
The present findings leave open the possibility that the spatial realignment is not performed by the auditory cortex, but that the auditory cortical activity reflects the processing outcomes of the subcortical pathway. The cues involved in the grouping and spatial realignment of the vowel stimulus are extracted early on in the auditory pathway. Horizontal localization cues are analyzed by the superior olive in the auditory brainstem (Grothe et al., 2010), implying that the location information associated to the individual components is available after this stage. There is also evidence for sound onset, which may have cued the fusion of our spatially separated vowels, being emphasized from the early stages of the auditory pathway, starting from the auditory nerve (Schnupp et al., 2011). A potential site for the spatial realignment could then be, for instance, the inferior colliculus, a major auditory nucleus in the midbrain that receives direct input from the superior olive and the cochlear nucleus. Assessing the details of subcortical processing is challenging in humans, but this hypothesis could possibly be tested by means of electrophysiology in experimental animals. The processing mechanisms described here are unlikely to be specific to humans as they contribute to the analysis of complex auditory scenes, a task that is not specific to human needs.

In natural listening situations, sounds that originate from the same source also tend to arise from the same location. Therefore, it seems surprising that the auditory nervous system has the ability to group components regardless of such separations, and even to modify spatial information in the process. There are, however, natural conditions under which sound energy related to the same object can arise from multiple locations, namely, when there are reflections. In such situations, localization is dominated by the first wavefront of the sound while the later arriving energy is interpreted as echoes (Litovsky et al., 1999). The echoes no longer contribute to location perception even though they are deemed to belong to the same object (Bradley and Souldore, 1995). This phenomenon, known as the precedence effect, provides another example of location information that gets suppressed in grouping although, as an important difference from the object formation associated to the individual components is line with this psychoacoustical result: the even-formant component was aligned toward the location of the odd-formant component. It therefore appears that the perceived location for the vowel object is dominated by the location information contained by the odd-formant component. While both components had a prominent interaural level difference (ILD), the dominance of the odd-formant component may have been due to this component containing the frequency range (<1.3 kHz) most useful for localization based on interaural time difference (ITD).

Previously, ITD has been found to dominate location perception whenever sounds contain low-frequency energy, even when other cues provide opposing evidence (Wightman and Kistler, 1992; Macpherson and Middlebrooks, 2002). Based on our present and previous findings (Takanen et al., 2013), the localization of the auditory object when fusing spatially separated components seems to follow similar principles.

In the vowel stimulus of the present study, both odd- and even-formant components had a prominent ILD. Therefore, the differences in location are unavoidably accompanied by differences in monaural sound level. For instance, a sound source in the right gives rise to a higher sound level in the right ear than a sound source in the left. Based on this, one might hypothesize that the stronger adaptation caused by adaptors in one spatial location than in another is simply due to sound level: for instance, the adaptor in the right location would cause stronger adaptation in the left hemisphere simply by virtue of higher sound level in the contralateral ear. This hypothesis is, however, not supported by the data measured. If the location-specific adaptation found here would be due to monaural sound level, it should manifest itself in both hemispheres and, further, should be independent of the probe sound. In contrast, the adaptation recorded here was strongly dependent both on cortical hemisphere and on the probe configuration: the location-specific adaptation occurred only in the hemisphere contralateral to the perceived probe location. This shows that the adaptation cannot be explained by monaural sound level, but instead reflects location-specific mechanisms.

In conclusion, the construction of auditory objects from spatially separated sound components relies on a neural reorganization of the auditory scene. In this process, the sound components are spatially realigned so that the resulting auditory object can be localized. This makes perceptual fusion regardless of a spatial conflict possible and may also be a neural mechanism that contributes to difficulties in the segregation of concurrent sound sources. Here, the neural correlates of this process were observed in the auditory cortex within the first 100 ms after the sound onset. This places the fusion between sound components and, further, the allocation of spatial locations to auditory objects into the auditory cortex or, perhaps, an even earlier stage of the auditory pathway. These findings show that the human cortex does not retain a faithful representation of the physical sound scene, but rather represents the scene as we perceive it.

ACKNOWLEDGMENTS

This work was supported by the Academy of Finland (Grant Nos. 13257844 and 13251770). The research leading


