Integration and segregation in auditory scene analysis

Elyse S. Sussman

Department of Neuroscience, Albert Einstein College of Medicine, 1410 Pelham Parkway South, Bronx, New York 10461

(Received 12 January 2004; revised 29 November 2004; accepted 10 December 2004)

Assessment of the neural correlates of auditory scene analysis, using an index of sound change detection that does not require the listener to attend to the sounds [a component of event-related brain potentials called the mismatch negativity (MMN)], has previously demonstrated that segregation processes can occur without attention focused on the sounds and that within-stream contextual factors influence how sound elements are integrated and represented in auditory memory. The current study investigated the relationship between the segregation and integration processes when they were called upon to function together. The pattern of MMN results showed that the integration of sound elements within a sound stream occurred after the segregation of sounds into independent streams and, further, that the individual streams were subject to contextual effects. These results are consistent with a view of auditory processing that suggests that the auditory scene is rapidly organized into distinct streams and the integration of sequential elements to perceptual units takes place on the already formed streams. This would allow for the flexibility required to identify changing within-stream sound patterns, needed to appreciate music or comprehend speech. © 2005 Acoustical Society of America. [DOI: 10.1121/1.1854312]

PACS numbers: 43.66.Ba, 43.66.Mk, 43.64.Bt [SEM]

Pages: 1285–1298

I. INTRODUCTION

From infancy we experience a complex auditory environment made up of several simultaneously active sound sources that often overlap in many acoustic parameters. Yet we experience a coherent auditory environment made up of identifiable auditory events. Auditory scene analysis (ASA) involves the ability to disentangle the mixture of sound input, integrating sensory inputs that belong together and segregating those inputs that originate from different sources (Bregman, 1990). Accordingly, segregation and integration are two fundamental aspects of ASA. The focus of the current paper is on the interaction between these two important auditory processes in ASA.

A. Auditory memory

A crucial part of the auditory scene analysis process is the ability to connect elements over time (integration processes), allowing us to recognize a series of footsteps or to understand spoken speech. Transient memory plays a critical role in this process especially because auditory processing relies so heavily on the temporal domain. This is easily demonstrated by thinking about how we understand a spoken sentence. Once each word of a sentence is spoken (i.e., the physical source of auditory information is completed), only the neural trace of the physical sound information is retained. To get to the meaning of the sentence we have to access the memory of the previous words that were spoken. Considering, however, that processing a speech stream most often occurs within the context of other simultaneous sound streams, whether on a city street, at the office, or in a department store. The focus of the current study is on understanding the relationship between the integration and segregation processes, specifically, to focus on the order in which the following processes are hypothesized to occur: (a) the segregation of input into distinct streams (perceived environmental sources) and then (b) the integration of elements necessary for making sense of the sequential patterns within a particular sound stream. In understanding how this memory operates in facilitating ASA, it is important to understand the relationship between these processes in terms of how auditory events are represented and stored in memory. Event-related brain potentials (ERPs) were used in the current investigation to explore the contents of transient auditory memory and to take advantage of the ability to use this method to observe the relationship between these processes when the listener has no task involving the sounds.

B. Event-related brain potentials

ERPs are noninvasive electrophysiological measures of cortical responses to sensory events that provide distinctive signatures for stimulus change. Change is an important cue for the auditory system. Because ERPs provide such high temporal resolution (in the order of milliseconds) and are time-locked to stimulus events, we can gain information about the timing of certain cognitive processes evoked by a given sound. Of particular importance for observing unattended sound processing is the mismatch negativity (MMN) component of ERPs ( Näätänen, 1992). MMN reflects sound change detection and is elicited even when the sounds have no relevance to ongoing behavior. The change detection process underlying MMN generation is dependent upon auditory memory. It uses sensory representations of the acoustic regularities extracted from the sound sequence (often called the “standard”). These sensory representations of the standard form the basis for the change detection process. Incoming
sounds that deviate from the neural trace of the standard elicit MMN, which is generated in the auditory cortex and is usually evoked within 200 ms of change detection. Thus, MMN represents an early process of change detection based upon a memory of the previous sound stimulation. An indication that the mechanisms underlying the MMN system involve memory processes is found in evidence that MMN generation relies on current flow through active N-methyl-D-aspartate (NMDA) channels (Javitt et al., 1996). Javitt et al. showed that blockage of NMDA, which has been associated with the cellular mechanisms involved in memory processes (Compt et al., 2000; Kauer et al., 1988; Pulvirenti, 1992), prevented MMN generation without disturbing elicitation of the prior obligatory cortical response to sound onsets. This suggests a link between working memory processes and the cortical auditory information-processing network involved in MMN elicitation.

The main neural generators of the MMN are located bilaterally in the supratemporal plane, as determined by dipole-modeling of electric (Scherg et al., 1989; Ha et al., 2003) and magnetic responses (Sams and Hari, 1991; Woldorff et al., 1998), scalp-current density maps of scalp-recorded ERPs (Giard et al., 1990), functional magnetic resonance imaging (fMRI; Opitz et al., 1999), and intracortical ERP recordings (Halgren et al., 1995; Kroppotov et al., 1995). The location of the generators in the auditory cortex accounts for the observed scalp topography of the waveform, which is maximally negative over the fronto-central scalp locations and inverts in polarity below the Sylvian fissure.

C. Assessing neural representations of sound organization

MMN can be used to probe the neural representations extracted from the ongoing sound input because the response to a particular sound is based upon the memory of the previous sounds. This is illustrated in Fig. 1. In the well-known auditory oddball paradigm, an “oddball” or infrequent sound is presented randomly among instances of a frequently repeating sound. The oddball elicits MMN because it is detected as deviating, in some sound feature (frequency, intensity, duration, spatial location), from the frequently repeating sound. In the top panel of Fig. 1, a regular auditory oddball paradigm is shown. The letter “X” represents a tone of one frequency and the letter “O” represents a tone of a different frequency. The oddball (O) would elicit MMN because it has a different frequency than that of the standard (X). In the bottom panel of Fig. 1, the same ratio of O to X sounds is presented, but instead of presenting the O tone randomly, it occurs every fifth tone in the sequence. If the brain detects the regularity, in this case the five-tone repeating pattern (X-X-X-X-O-X-X-X-X-O-...), then no MMN will be elicited by the O tone (the “oddball” in the randomized context) because the O tone is part of the standard repeating (five-tone) regularity. It is not deviant. Thus, it is important to note that MMN is not ipso facto elicited when there is a situation of a frequent and an infrequent tone presented in a sound sequence. What is crucial to its elicitation is what is detected as the regularity in the sound sequence and stored in auditory memory. Thus, MMN is highly dependent upon the context of the stimuli, whether that context is detected without attention focused on the sounds (Sussman et al., 1999a, 1998a) or whether the context is influenced by attentional control (Sussman et al., 1998b; 2002a).

Thus, the change detection process is far more complex than is superficially evident from the auditory oddball paradigm, the most common paradigm used to elicit MMN. Simple sensory discrimination of two tones (i.e., being able to detect whether one tone is different from another in any sound feature) requires only a single presentation of each sound. Repetition is not necessary for this type of discrimination. However, the MMN process is based upon repetition and the detection of the regularities in the acoustic input. Even with the seemingly simple auditory oddball paradigm, deviance detection is based upon the memory of the larger context of the sounds and not simply on the detection of two different tones in the sound sequence. Specifically because MMN relies on the larger auditory context (sequential processing of auditory events), it can be used to assess how the acoustic information is structured and stored in memory.

D. Segregation of sounds to create distinct sources

Segregation of sounds to distinct auditory sources is a key function of the auditory system. Consider that the acoustic information entering one’s ears is a mixture of all the sounds in the environment, without separation. Decomposing the auditory input is therefore a crucial step, allowing us to detect a single voice in a crowd or to distinguish a voice coming from the left or right in the proverbial “cocktail party” setting. For that reason, the process of disentangling the sound into distinct sources plays a critical role in how we experience the auditory environment. Although there remains controversy over the role of attention in the stream segregation process (Botte et al., 1997; Bregman, 1990; Brochard et al., 1999; Carlyon et al., 2001; Jones et al., 1978; 1999; Macken et al., 2003; Sussman et al., 1999a; Winkler et al., 2003), there is considerable ERP evidence to suggest that the segregation of auditory input to distinct sound streams can occur without attention focused on the sounds, and that auditory memory can hold information about multiple sound streams independently (Sussman et al., in press, submitted, 1999a; Ritter et al., 2000; Winkler et al., 2003). This basic finding has been replicated in various paradigms using dif-
ferent segregation cues, such as frequency proximity (Sussman et al., 2001; Winkler et al., 2003), temporal proximity (Sussman et al., 1999a), and a combination of spatial location, frequency, and intensity (Ritter et al., 2000). These studies provide evidence in humans that auditory stream segregation mechanisms are part of an early, primitive process. This is consistent with evidence from animal studies demonstrating that the basic stream segregation mechanisms are part of vertebrates’ auditory systems (Fay, 2000; Hulse et al., 1997; Fishman et al., 2001). It is functionally parsimonious that the basic stream segregation process would occur automatically (calculated on the basis of the acoustic characteristics of the sensory input), so that attention, a limited human resource, could be used to identify patterns within the already organized streams, needed for understanding speech or listening to music.

E. Integration of sound elements within a single stream

Auditory integrative processes have predominantly focused on how successive elements influence a single percept of sound, such as with the phenomena of loudness summation (Zwisloski, 1969) and auditory recognition masking (Massaro, 1975). The perception of a sound event is often determined by the sounds that surround it. Recently, however, it was demonstrated that the larger context of a sequence of sounds could affect processing of the individual sound components even when not in direct temporal proximity (Sussman et al., 2002b; Sussman and Winkler, 2001). It is this context-dependent integration process that is the focus of the current study and described in detail below.

The paradigm of the current study was based on previous work (Sussman et al., 1999b, 2002b; Sussman and Winkler, 2001) showing that two different deviations from the same repetitive standard, occurring successively within a less than 200-ms interval (called a “double deviant”), elicit only one MMN. However, when double deviants were presented within a stimulus sequence that also contained single deviants (i.e., sounds that differ from the repetitive standard in one of the two ways of the double deviants), the double deviants elicited two discrete MMNs, which were separated by the temporal distance of the successive deviations (a schematic of the basic paradigm used in these studies is shown in Fig. 2). Initially, we hypothesized that the successive deviants were integrated into a single deviant event because they were indistinguishable as discrete events when subjects ignored the sounds. This led to a prediction that if subjects could actively discriminate the two successive deviants in the blocked context, then the double deviants would elicit two MMNs. However, this turned out not to be the case. We found that even when subjects were trained to discriminate the two successive deviants, only one MMN was still elicited by them (Sussman et al., 2002b). The integration phenomenon occurred solely as a function of changing the stimulus-driven context (i.e., blocked versus mixed) irrespective of whether attention was focused on the sounds.

The results of these studies are consistent with a context-based interpretation of auditory event formation in that the relevant information influencing event formation can be calculated over a longer time. In the blocked context, the second deviant is fully predicted by the first. Every time a deviant occurs another deviant follows it (termed “double deviants”). In the mixed presentation, both single and double deviants occur randomly in the sequence. See text for further discussion.
II. EXPERIMENT 1. SEGREGATION AND INTEGRATION (LOW STREAM)

Four conditions were presented in experiment 1, in which the context was manipulated within the low tones stream. In one condition the low stream of sounds had only double deviants (alternating-blocked) and in another condition double deviants were presented along with single deviants (alternating-mixed). Two control conditions were used to ascertain the effect of an intervening sound within a single stream (low only—standard intervening condition) and to determine whether the integration of elements was initiated by a time constant or by the number of elements occurring successively (low only—triple deviant condition).

A. Participants

Thirteen adults (five males, eight females) between the ages of 23 and 42 years, with no reported hearing loss or neurological disorders, were paid to participate in the study. All subjects passed a hearing screening. Participants were informed of the experimental protocol before they signed a consent form. The study was conducted at the Albert Einstein College of Medicine.

B. Stimuli

Three tones (one high frequency and two low frequency) were used to create the stimulus sequences. They were 50 ms in duration with an intensity of 80 dB SPL (calibrated with Bruel & Kjaer 2209 sound-level meter). The high-frequency tone was 1568 Hz (H) and the low-frequency tones were 440 Hz (L) and 494 Hz (LD). Onset and offset ramps of 7.5 ms were created using a Hanning window for all tones. Stimuli were presented in four conditions: (a) alternating-blocked, (b) alternating-mixed, (c) low only—standard intervening, and (d) low only—triple deviant. In the two alternating conditions, H and L alternated in a regular fashion at a rate of 75 ms SOA (onset to onset), except when the deviant (LD) occurred. Thus, a low tone (or a high tone) occurred every 150 ms. In the alternating blocked condition, the deviant (LD) always occurred twice in succession in the low stream (e.g., H L H L H L H LD H LD H L H L L); these will be called “double deviants” (see Fig. 3). Double deviants occurred randomly 7.5% of the time. In the alternating mixed condition, LD was randomly presented both singly (these will be called “single deviants”) 5% and twice in succession 5% each (e.g., H L H L H L H LD H LD H L H L H LD H LD H L; see Fig. 3). LD was presented 15% of the time overall in both conditions. Thus, the “blocked” and “mixed” nomenclature refers to the low stream only, specifically to the deviants and whether there are only double deviants in the sequence (blocked) or whether there are mixed single deviants with double deviants (mixed) in the same sequence.

C. Procedures

Participants sat in a comfortable chair approximately 1 m from a TV monitor in a sound-attenuated booth. They were instructed to watch a closed-captioned video, itself containing no sounds, and to ignore the sounds. Three thousand stimuli were presented in each of the four conditions (3000 × 4 = 12,000 stimuli in all). There were two runs of 1500 stimuli presented, totaling eight runs. The eight runs were presented randomly, counterbalanced across subjects to avoid any possible order effects.

D. Electroencephalogram (EEG) recording and data analysis

EEG was recorded with an electrode cap using the following scalp locations (10–20 system): Fz, Cz, Pz, F3, F4, C3, C4, and both mastoids (LM and RM, for the left and right mastoids, respectively). Horizontal eye movements were measured by recording the horizontal electro-oculogram (HEOG) between electrodes F7 and F8 using a bipolar montage and vertical using a bipolar montage between FP1 and an external electrode placed beneath the left eye (VEOG). The common reference electrode was attached to the tip of the nose. Eye movement was monitored with VEOG and HEOG. HEOG was monitored for eye saccade
movement to ensure that participants were reading the captions on the video. The EEG was digitized (Nicolet amplifiers) at a sampling rate of 250 Hz (bandpass 0.05–100 Hz) and then off-line filtered between 1 and 30 Hz. Epochs of 600 ms, with a 100-ms prestimulus and 500-ms poststimulus period, were averaged separately for the standards and double deviants. The double deviants (and triple deviant) were averaged from the onset of the first of the successive deviants, thus the response to the successive deviants could be observed within this period. Epochs with an electrical change exceeding 100 μV were excluded from further averaging to remove the trials contaminated by artifacts of non-cortical origin.

For each participant, the artifact-free epochs were then averaged separately for the four conditions and two stimulus types (standard and deviant). The mean amplitude in the 100-ms prestimulus period was subtracted from each point of the averaged ERP responses, which served as the reference from which the evoked responses were measured.

The difference waveforms were obtained by subtracting the ERP response to the standard (e.g., L-H-L in the alternating condition) from the ERP response to the deviants. In the mixed conditions, double deviants (LD-H-LD) were averaged separately from single deviants (LD-H-H). The intervals used for statistically evaluating for the presence of the potential MMNs were determined from the group mean deviant-minus-standard waveforms at Fz (the site of greatest signal to noise ratio), taking a 40-ms window centered on the peak of the MMNs from the conditions in which two successive MMNs were elicited (the low only—standard intervening and alternating-mixed conditions). The mean amplitude obtained in the intervals used for the two low-only conditions (range 1: 116–156 ms and range 2: 304–344 ms) and for the two alternating conditions (range 1: 140–180 ms and range 2: 320–360 ms) were measured for each subject individually and were then used to verify the presence of MMN, for each range and condition separately with one-sample, one-tailed t-tests using the data from the Fz electrode site. Analysis of variance (ANOVA) for repeated measures with factors of condition (blocked versus mixed) and range (range 1 versus range 2) was used to compare the mean voltage. Greenhouse–Geisser correction was reported when appropriate. Tukey HSD posthoc tests were calculated.

E. Results and discussion

The ERP results for the low-only and alternating conditions are displayed in Figs. 4 and 5, respectively (also see Table I).

1. Low-only conditions

In the standard-intervening condition, the two deviants elicited separate MMNs ($t_{12}=6.14$, $p<0.01$, $t_{12}=2.11$, $p<0.03$; ranges 1 and 2, respectively) approximately 150 ms from each other (see Fig. 4, top). This demonstrates that the two deviants were processed as separate events when there was just one standard tone intervening between them within the single stream. In the triple-deviant condition, the triple-deviant elicited only one MMN ($t_{12}=6.87$, $p<0.01$, $t_{12}<1$, $p>0.72$; ranges 1 and 2, respectively; see Fig. 4, bottom), suggesting that event formation within the integrating period operates independently of the number of elements falling within the window. The single presentation of a standard tone between the two deviants (e.g., LD-L-LD) resulted in the deviants being processed as two separate events in the standard-intervening condition, compared to the triple-deviant condition in which all three of the elements presented in the same time window were processed as a single deviant event (LD-LD-LD). This provides evidence that event formation operates on the basis of the type of elements falling within the integrating period, not simply integrating...
all elements by a time constant. The integration or segregation of successive elements within the low stream was governed by the nature of the information: elements carrying the same information were integrated into a single auditory event, whereas elements that differed within that same time period were distinguished from each other. This suggests that the underlying time constant that tends to integrate information operates on the basis of the distinctiveness of the information in the larger context.

2. Alternating conditions

In the blocked condition, double deviants in the low stream elicited one MMN ($t_{12} = 2.52, p < 0.02, t_{12} < 1, p > 0.50$; ranges 1 and 2, respectively; see Fig. 5, top row). The double deviants were integrated into a single deviant event even though there was a tone intervening between them. This indicates that the streams were segregated prior to the integration process. In contrast, in the mixed condition, the same double deviant events elicited two MMNs ($t_{12} = 2.33, p < 0.02, t_{12} = 5.05, p < 0.01$; ranges 1 and 2, respectively; Fig. 5, bottom row) when single deviants also occurred within the low stream. ANOVA comparing the mean voltages of the difference waveforms in ranges 1 and 2 revealed an interaction between condition (blocked versus mixed) and range [$F(1, 12) = 6.48, p > 0.03$]. Posthoc analyses showed that this interaction was due to larger mean amplitude for range 1 than range 2 in the blocked condition and no difference between mean amplitude in ranges 1 and 2 of the mixed condition. This confirms that one MMN was elicited in the blocked condition and two MMNs were elicited in the mixed condition. This result demonstrates that the presence of the single deviants in the low stream altered the MMN response to the double deviants in the same way as occurs when no intervening high tones are presented in the stimulus block [i.e., when the low stream is presented alone at the same stimulus rate (e.g., Sussman et al., 2002b)].

The double deviants in the alternating-blocked condition elicited one MMN but they elicited two MMNs in the alternating-mixed condition. Double deviants in both alternating conditions should have elicited two MMNs if the successive deviant-standard-deviant pattern had been processed as a single stream, as demonstrated by results of the low-only conditions. However, the presence of the single deviants within the low stream of the mixed condition altered the MMN response to the double deviants. This difference in the MMN response to the same double deviants for the blocked condition and the alternating-blocked condition suggests that one MMN was elicited in the blocked condition but two MMNs were elicited in the alternating-blocked condition. This result demonstrates that the presence of the single deviants in the low stream altered the MMN response to the double deviants in the same way as occurs when no intervening high tones are presented in the stimulus block [i.e., when the low stream is presented alone at the same stimulus rate (e.g., Sussman et al., 2002b)].

III. EXPERIMENT 2. SEGREGATION AND INTEGRATION IN BOTH HIGH AND LOW STREAMS

A. Rationale

The within-stream contextual effects observed in experiment 1 occurred within only one of the two possible streams. This leaves open the question of whether integration effects occurred within-stream because there was only one stream with changing information. Would the complexity (having
deviants in both streams) created by simultaneously changing contexts within the high and low streams may modify processing of the sounds? If complexity could negate the integration effects mediated by changing the contextual information, then a global effect on sound processing could account for the results and could not be excluded as an interpretation of the results of experiment 1. If, however, the same pattern of integration effects were found for double deviants occurring within both the low and high streams (double deviants eliciting one MMN in blocked and two MMNs in mixed contexts) then the conclusions of experiment 1 would be expanded and strengthened. To test this, the same basic paradigm of experiment 1 was used except that deviants occurred in both streams.

B. Participants

Twelve adults (five males, seven females) between the ages of 22 and 37 years, with no reported hearing loss or neurological disorders, were paid to participate in the study. All subjects passed a hearing screening. Participants were informed of the experimental protocol before they signed a consent form. The study was conducted at the Albert Einstein College of Medicine. Two subjects’ data were removed due to excessive eye artifact. One subject’s data were removed due to experimental error. The data from the remaining nine subjects are reported. None of the subjects that participated due to experimental error. The data from the remaining nine subjects were removed. One subject’s data were removed due to excessive eye artifact. One subject’s data were removed due to experimental error. The data from the remaining nine subjects are reported. None of the subjects that participated in experiment 2 participated in experiment 1.

C. Stimuli

Four tones (two high frequency and two low frequency), 50 ms in duration (tone intensity was calibrated with Bruel & Kjær 2209 sound-level meter at a level of 80 dB SPL), were used to create the stimulus sequences. The high-frequency tones were 1568 Hz (H) and 1760 Hz (HD) and the low-frequency tones were 440 Hz (L) and 494 Hz (LD). Onset and offset ramps of 7.5 ms were created with a Hanning window for all tones. Stimuli were presented in two conditions: (a) alternating-blocked and (b) alternating-mixed. H and L alternated in a regular fashion at a rate of 75 ms SOA (onset-onset). Thus, a low tone (or a high tone) occurred once every 150 ms. In the alternating-blocked conditions, double deviants (LD or HD) always occurred twice in succession (presented randomly 6% of the time) in their respective stream but deviants never overlapped across streams. In the alternating-mixed conditions, HD and LD were presented randomly as single deviants (4%) and twice in succession (4%) each. Thus, HD and LD were presented 12% overall in both conditions. The “blocked” and “mixed” nomenclature refers to the presentation of the deviants in both the high- and low-frequency streams.

D. Procedures

Participants sat in a comfortable chair approximately 1 m from a TV monitor in a sound-attenuated booth. They were instructed to watch a closed-captioned video, itself containing no sounds, and to pay no attention to the sounds. Six thousand stimuli were presented in the blocked condition in two runs of 3000 stimuli, and 15,000 stimuli in five runs of 3000 stimuli were presented in the alternating condition, totaling seven runs. The seven runs were presented randomly, counterbalanced across subjects to avoid any possible order effects.

E. Electroencephalogram (EEG) recording and data analysis

Recording parameters matched those of experiment 1 with the following exceptions. The EEG was digitized (Neuroscan Synamps amplifiers) at a sampling rate of 500 Hz (bandpass 0.05–100 Hz) and then off-line filtered between 1 and 15 Hz. Epochs with an electrical change from baseline exceeding ±75 μV were excluded from further averaging.

For each participant, the artifact-free epochs were then averaged separately for two conditions (blocked and mixed), two frequency ranges (high and low), and two stimulus types (standards and double deviants).

The difference waveforms were obtained by subtracting the ERP response to the standard from the ERP response to the deviant. The intervals used for statistically evaluating for the presence of the potential MMNs were determined from the group mean deviant-minus-standard waveforms at Fz (the site of greatest signal-to-noise ratio), taking a 40-ms window centered on the peak of the MMNs for range 1 and range 2, which can be observed in the epoch. Range 2 intervals for the blocked conditions were obtained from the corresponding mixed conditions, in which two MMNs were elicited. The mean amplitude obtained in the intervals used for the blocked—high condition was range 1: 106–146 ms and range 2: 270–310 ms; for the blocked—low condition was Range 1: 126–166 ms and range 2: 304–344 ms; for the mixed—high condition was range 1: 100–140 ms and range 2: 270–310 ms; and for the mixed—low condition was range 1: 126–166 ms and range 2: 304–344 ms. The intervals were measured for each subject individually and were then used to verify the presence of MMN. Because the predictions for the MMNs were a priori, one-sample one-tailed Student’s t-tests were used to determine whether the mean amplitudes obtained in these latency ranges were significantly greater than zero, separately for each range and condition. ANOVA for repeated measures with factors of condition (blocked versus mixed), frequency (high versus low) and range (range 1 versus range 2) was used to compare mean voltage. Greenhouse–Geisser correction was reported when appropriate. Tukey HSD posthoc tests were calculated.

F. Results and discussion

The ERP results for the alternating-blocked and alternating-mixed conditions are displayed in Figs. 6 and 7, respectively (also see Table II).

1. Alternating-blocked condition

Double deviants elicited one MMN in the low stream ($t_9=5.1$, $p<0.01$, $t_9<1$, $p>0.43$; ranges 1 and 2, respectively; see Fig. 6, top row) and one MMN in the high stream ($t_9=2.0$, $p<0.05$, $t_9=1.0$, $p>0.16$; ranges 1 and 2, respectively; see Fig. 7, top rows). Consistent with the results of experiment 1, blocked condition double deviants elicited one MMN.
condition altered the MMN response to the double deviants, as compared to the blocked conditions, independently in each stream.

The results of experiment 2 rule out complexity as an explanation for the within-stream integration effects found in experiment 1, and extend the results of experiment 1 to show that integration processes are carried out concurrently on multiple streams of auditory information.

IV. EXPERIMENT 3. SEGREGATION BY NONSPECTRAL CUE

A. Rationale

Experiments 1 and 2 used pure tone stimuli with a large (15 semitone) difference between high and low sounds. It is therefore possible to consider that the integration effects may have taken place by frequency range, and were not mediated by streaming. An alternative interpretation would be that with no overlapping excitation of hair cells within the cochlea, frequency effects rather than segregation effects may explain the integration effects of experiments 1 and 2. This interpretation of the current results would characterize frequency as the dominant factor influencing the integration processes, independent of segregation. Thus, another experiment was conducted to assess the integration effect when stream segregation was cued by a factor other than spectral information. The same basic paradigm of experiment 1 was used, except that the alternating sequence of sounds included complex tones and narrowband noise bursts with overlapping spectral composition. The noise and tones were expected to segregate to separate streams (Singh and Bregman, 1997; Iverson, 1995; Cusack and Roberts, 1999, 2000). If the integration were a frequency-specific phenomenon in the previous two experiments, that is, if the integration process did not depend on the initial segregation of the information, then two MMNs should be elicited in both blocked and mixed contexts, because of the effect of the intervening tones (see experiment 1, discussion of low-only control conditions). This result may be consistent with a peripheral channeling explanation (Beauvois and Meddis, 1991; Hartmann and Johnson, 1991; Rose and Moore, 2000). On the other hand, if a similar pattern of MMN results occurs as did in experiments 1 and 2, then it could be concluded that the segregation process preceded the integration process. Or, stated differently, we can conclude that the integration of sequential sound elements takes place within the distinct sound streams.

B. Participants

Twelve adults (five males, seven females) between the ages of 22 and 37 years, with no reported hearing loss or neurological disorders, were paid to participate in the study. All subjects passed a hearing screening. Participants were informed of the experimental protocol before they signed a consent form. The study was conducted at the Albert Einstein College of Medicine. Two subjects’ data were removed due to excessive eye artifact. One subject’s data were removed due to experimental error. The data from the remaining nine
C. Stimuli

Two types of complex sounds were used (created with Adobe Audition 1.0 software): harmonic complexes (called “tones”) and bandpass-filtered noise bursts (called “noise”). The harmonic complexes had the same fundamental frequency (f0) as the low tones of experiments 1 and 2 (440 Hz for the standard sound and 494 Hz for the deviant sound) and were composed of the first three harmonics with equal amplitudes. Tone 1 (T) consisted of 440, 880, and 1320 Hz, and tone 2 (TD) consisted of 494, 988, and 1482 Hz. The other type of complex tone (N) was created by narrow-band fast Fourier transform (FFT) filter performed on broadband noise between 440 and 1320 Hz. Thus, it overlapped the same spectral region (440–1320) as the standard complex tone (T). Tone duration for all tones was 50 ms, with onset and offset ramps of 7.5 ms made with a Hanning window. Sound intensity of the tones and noise were equated and presented at 77 dB SPL (Briel & Kjaer 2209 sound-level meter).

Stimuli were presented in four conditions: tones-only blocked, tones-only mixed, alternating-blocked and alternating-mixed. The tones-only conditions were control conditions to verify that the auditory context would have a similar effect on complex tones as they did with the pure-tone stimuli in previous studies (i.e., double deviants eliciting one MMN in blocked context and two MMNs in mixed context). In the event that the effects of context on integration of sequential sounds were specific to pure tones, the control conditions were needed to determine this. The sound sequence was presented at 150 ms onset-to-onset pace, the pace that the complex tones would occur in the alternating conditions (alternating-blocked and alternating-mixed) of the main experiment. Three thousand stimuli were presented in two runs of 1500 stimuli each, for the tone-only blocked condition and 6000 stimuli in four runs of 1500 stimuli for the tone-only mixed condition.

In the alternating conditions, tone and noise stimuli were alternated (T-N-T-N-T-N-···) at an onset-to-onset pace of 75 ms. The alternating conditions used the same basic paradigm as experiment 1, with the context changes occurring only in the tones stream. Six thousand stimuli were presented in two runs of 3000 stimuli for the alternating-blocked condition and 24000 stimuli in four runs for the alternating-mixed condition. Deviants (TD) occurred 15% overall in all conditions (7.5% occurrence of double deviants in the blocked conditions and 8% single deviants and 7% double deviants in the mixed conditions).

D. Procedures

Participants sat in a comfortable chair approximately 1 m from a TV monitor in a sound attenuated booth. They were instructed to watch a silent closed-captioned video and to pay no attention to the sounds. Three thousand stimuli were presented in two runs of 1500 stimuli each, for the tone-only blocked condition and 6000 stimuli in four runs of 1500 stimuli for the tone-only mixed condition. For the alternating conditions, 6000 stimuli were presented in two runs of 3000 stimuli for the alternating-blocked condition and 24000 stimuli in four runs for the alternating-mixed condition. Deviants occurred 15% overall in all conditions (7.5% occurrence of double deviants in the blocked conditions and 8% single deviants and 7% double deviants in the mixed conditions).
conditions. The order of the four conditions was counterbalanced across subjects to avoid any possible order effects.

E. Electroencephalogram (EEG) recording and data analysis

Data recording parameters were the same as those used in experiment 2. For each participant, the artifact-free epochs were averaged separately for the complex tones by stimulus type [standards (T) and double deviants (TD)], separately, in the four conditions (tones-only blocked, tones-only mixed, alternating-blocked, and alternating-mixed).

The difference waveforms were obtained by subtracting the ERP response to the standard from the ERP response to the deviant. The intervals used for statistically evaluating for the presence of the potential MMNs were determined from the group mean deviant-minus-standard waveforms at Fz (the site of greatest signal-to-noise ratio), taking a 40-ms window centered on the peak of the MMNs where elicited. For the blocked conditions, range 2 intervals from the mixed conditions were used to measure the presence of MMN in range 2 of the blocked conditions because MMN was expected to be elicited in the mixed conditions. For the tones-only blocked condition the mean amplitude was obtained in the following intervals, range 1: 104–144 ms and range 2: 282–322 ms; for the tones-only mixed condition the intervals were for range 1: 100–140 ms and range 2: 282–322 ms; for the alternating-blocked condition the intervals were for range 1: 104–144 ms and range 2: 258–298; and for the alternating-mixed condition the intervals were for range 1: 100–140 ms and range 2: 258–298. Intervals were measured for each subject individually and then used to verify the presence of the MMN using a repeated measures ANOVA with factors of condition (alternating-blocked, tones-only blocked, alternating-mixed, tones-only mixed) × peak (range 1 versus range 2) and stimulus type (deviant versus standard) to determine in which conditions and which peaks the standard and deviant waveforms differed from each. ANOVA with factors of condition (blocked versus mixed), control (alternating versus tones-only), and range (range 1 versus range 2) was calculated on the difference waveforms to compare mean voltages. Greenhouse–Geisser correction was reported when appropriate. Tukey HSD post hoc tests were calculated.

F. Results and discussion

The ERP results for the tones-only and alternating conditions are displayed in Figs. 8 and 9, respectively (also see Table III).

The ANOVA revealed a three-way interaction between condition, stimulus type, and range \( [F(3,24) = 6.25, p < 0.01] \), showing that the significant difference between the standard and deviant (MMN) was dependent upon the condition and range. Tukey HSD post hoc comparisons revealed that MMNs were elicited in both range 1 and range 2 of the mixed conditions [alternating mixed and tones-only mixed; see Figs. 8 (bottom row) and 9 (bottom row)], whereas MMNs were elicited only in range 1 of the blocked conditions [alternating blocked and tones-only blocked; see Figs. 8 (top row) and 9 (top row)]. This pattern of results is consistent with experiments 1 and 2, in which two MMNs were elicited by double deviants in the mixed context and one MMN was elicited in the blocked context.

The ANOVA comparing mean voltage of the deviant-minus-standard difference waveforms revealed an interaction between condition and range with no other interactions \( [F(1,8) = 21.11, p < 0.002] \). Post hoc analyses show that the mean voltage of range 1 was larger than that of range 2 for the blocked conditions and there was no difference in mean voltage between ranges 1 and 2 in the mixed conditions. This substantiates the finding that one MMN was elicited in the blocked conditions and two MMNs were elicited in the mixed conditions.
The results extend the findings of our previous studies using pure tones (e.g., Sussman et al., 2002b), now demonstrating with complex tones the same contextual effects on within-stream integration processes observed with pure tones. Additionally, the results demonstrate stream segregation by a nonspectral cue. This suggests that the stream segregation process involves more than simple frequency resolution of peripheral mechanisms. This finding is consistent with other studies showing that stream segregation is not solely dependent upon peripheral channeling (Bey and McAdams, 2003 [Experiment 1]; Cusack and Roberts, 2000; Iverson, 1995; Moore and Glockel, 2002; Roberts et al., 2002; Vliegen and Oxenham, 1999). The results provide strong support for the notion that segregation occurs prior to integration and that integration effects operate within already segregated sound streams. The interplay between these two processes is thus observed in complex sound situations, such as would be needed for processing speech and other environmental sounds.

V. GENERAL DISCUSSION

Three experiments were conducted to investigate the timing of two organizational processes of ASA, segregation of streams and integration of within-stream elements, when they were called upon to function concurrently and the listener had no task to perform with the sounds. The within-stream contextual effects, the integration of successive elements to a single or double deviant event as indexed by MMN, followed the same pattern of results as has been demonstrated for temporal integration processes when only one sound stream was present (Sussman et al., 2002b). This indicates that the segregation to independent streams occurred prior to the integration and formation of the auditory units within streams. The results demonstrate that (1) multiple sound streams are maintained independently, (2) stream segregation occurs prior to within-stream integration or event formation, and (3) auditory context effects act on the already formed streams.

Experiments 1 and 2 demonstrated that the integration process operates on two streams independently, as though each stream occurred alone. These results show that the integration occurred on the already segregated streams, the segregation process taking place at an earlier level of processing than the integration process. Experiment 3 confirmed that the segregation process preceded the integration of information within streams. Given that the spectral compositions of the two sets of sounds were similar in experiment 3, if the sounds were processed as part of the same stream, then the effects of intervening tones on within stream processes would have been observed (experiment 1, control conditions). Thus, the results of experiment 3 provided strong support to conclude that the segregation processes mediated the integration processes.

The timing of the segregation and integration processes that have been demonstrated in the current and previous studies suggests the following model. A neural representation of the sound input is initially made regarding how many sources are present in the signal. This is calculated according to the acoustic characteristics of the input (the frequency, intensity, duration, and spatial location features as well as the timing of the input). Extraction of the acoustic regularities is then established on the already formed streams (i.e., the context is determined). The long-term effects of the auditory context mediate local, within-stream, event formation (Sussman and Winkler, 2001). The integration processes (or sound event formation) operate after the within-stream context has

<table>
<thead>
<tr>
<th>Table III. Experiment 3. Mean amplitude values (in $\mu$V) and standard deviations (in parentheses) of the grand-averaged deviant-minus-standard (difference) waveforms in ranges 1 and 2 for all conditions at all electrode sites.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Condition</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Electrode</td>
</tr>
<tr>
<td>Fz</td>
</tr>
<tr>
<td>Cz</td>
</tr>
<tr>
<td>Pz</td>
</tr>
<tr>
<td>F3</td>
</tr>
<tr>
<td>F4</td>
</tr>
<tr>
<td>C3</td>
</tr>
<tr>
<td>C4</td>
</tr>
<tr>
<td>LM</td>
</tr>
<tr>
<td>RM</td>
</tr>
</tbody>
</table>

The context mediate local, within-stream, event formation, and that integration effects operate within already segregated sound streams. The interplay between these two processes is thus observed in complex sound situations, such as would be needed for processing speech and other environmental sounds.
been determined. Thus, whether a sound element is processed as one or two events depends on the surrounding information within the same stream. The MMN process uses this information—the representation of the segregated streams and the detected sound regularities within streams—in determining what is deviant in the sound input.

A. Implications for speech processing

Extending the notion that the integration of sound elements into perceptual units proceeds on the already segregated information, the data discussed here support the view that sound elements are integrated into linguistic units (phonemes, syllables, and words) after the initial segregation of the input into distinct sound sources. Speech processing, according to this model, would rely, at least in part, on the primitive ASA processes (Brown and Cooke, 1994; Rosenthal and Okuno, 1998).

When more than one voice occurs simultaneously, segregation by acoustic parameters (e.g., f0) serves to distinguish the separate voices in the auditory environment, whereas the integration of successive sound elements within the separate voices serves to distinguish speech units. There are additional segmentation processes that are also necessary for within-stream comprehension, those that allow us to distinguish words within a speech stream or individual footsteps within a series. For example, hearing the difference between “topic” and “top pick” is determined by the duration of the silent /p/ closure [less than 150 ms was needed to hear one word and between 150 and 300 ms was needed to hear two words; Pickett and Decker (1960)]. In fact, this timing that made the difference between hearing one or two words is consistent with our finding in which the same timing made the difference between tones (pure or complex) being represented as one or two deviant events. Two successive deviants elicited one MMN at 150-ms interval between them, whereas the two successive deviants elicited two MMNs when they were separated by 300 ms (Sussman et al., 1999b). Further research would be needed, however, to clarify whether pure acoustic cues function similarly as acoustic cues embedded in speech for defining auditory events.

Within-stream segmentation also takes place within the “integrating” period. The standard intervening tone between the two successive deviant tones (in the low-only—standard intervening condition) influenced the integration process as reflected by the MMN response (i.e., affected what became integrated into one event). When a single tone intervened between the two successive deviant tones, falling within the less than 200-ms integrating window, each of the deviants elicited its own MMN. What otherwise might have been responded to as one deviant event (two successive deviants occurring within 150 ms) was separated into two deviant events (eliciting two separate MMNs) by the presence of the intervening tone occurring between them when they were processed in the same low-tone stream. This result indicates that the integration of elements occurring within less than 200 ms operates according to the sequential organization of the unit or auditory event that falls within the window. The temporal proximity of the elements, on its own, is not enough for integration—this should already be evident because the phenomenon of streaming can occur when sounds are 100 ms apart—but neither is the frequency proximity sufficient (the low tone standards are near in frequency to the deviants but are not integrated just by the fact that they fall within this window). Thus, the result indicates that the integration mechanism is “searching” for elements that likely form a perceptual unit. The auditory system operates on the basis of the informational units falling within the window and does not just integrate all the elements that fall within a window of less than 200 ms.

This is further confirmed with the triple deviants in the low-only—triple deviants condition, in which three successive deviants occurring within the integrating period elicited one MMN. Elements that fall within the integrating window are being integrated to form a single auditory event. These data lend credence to the notion that the integrating period operates independently of the sheer number of elements falling within the window and focuses on the formation of events out of closely occurring sound elements. The integration of elements across time occurred as a function of the role of the stimuli that fell within the integrating period (e.g., deviant vs. standard). Whereas three deviant sounds occurring successively were responded to as a single deviant event, the successive deviant-standard-deviant pattern was not integrated, even though the tones occurred within the same time period. This demonstrates that auditory context influences the integrating window in terms of auditory event formation. This mechanism could be important in speech processing when determining which phonemic units are to be integrated and which segregated, such as those cues that distinguish “topic” from “top pick.”

It remains to be determined whether the segmentation processes that operate within streams to facilitate identification of perceptual units (and which are also part of integration processes necessary to link acoustic elements to perceptual units) differ from the procedures that are initially used to determine how many sources are in the environment. Additionally, the role of attention in these processes remains to be determined.

Context effects on MMN have now been demonstrated in a number of different studies using a wide range of paradigms (Sussman et al., 1998a, 1999a, 2001, 2002a, b; Sussman and Winkler, 2001; Winkler et al., 2003). Interestingly, MMN has also been reported to reflect categorical perception of consonants and vowels (Dehaene-Lambertz et al., 2000; Näätänen et al., 1997; Winkler et al., 1999), voice onset time (Sharma and Dorman, 2000), and length (Nenonen et al., 2003). The phonetic context effects (the shift in category boundary that is altered by surrounding speech characteristics) observed with these MMN studies of categorical perception might actually represent another form of context effect on MMN and not a speech-specific effect (Pisoni et al., 1983; Sussman et al., 2004).

B. Concluding remarks

The results of the current study demonstrate context effects on auditory event formation that occur after the segregation of sounds to distinct streams when the listener is not focused on the sound input. Moreover, integration of within-
stream elements is influenced by the larger context of the within-stream sounds. The auditory system does not just integrate all information that falls within a 150-ms window; the informational content of the elements determines what is integrated within this period and what is not. These data, taken together with previous results, also show that segregation occurs prior to other processes (e.g., Sussman et al., 1998b; 1999b; Yabe et al., 2001), provide strong evidence that segregation is an earlier, primitive process initially driven by stimulus characteristics of the acoustic input. Functionally, the integration of sequential elements to perceptual units taking place on the already segregated streams would facilitate the ability to identify within-stream sound patterns, thus making it possible to appreciate music or comprehend speech.

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health (Grant No. R01 DC04263). The author is grateful to Dr. Albert S. Bregman for his advice on experiment 3 and for his comments on an earlier version of the manuscript. The author thanks Dr. Yonathan Fishman and three anonymous reviewers for their helpful comments on an earlier version of the manuscript. The author is also grateful to Rosie Wong and Fakhar Kahn for collecting the data and for their assistance with data processing and manuscript preparation.

1This transient form of auditory memory has been estimated to store information for a period of time of at least 30 s (Cowan, 2001).


