Perceiving sequential dependencies in auditory streams^{a)}

Gerald Kidd, Jr.,^{b)} Christine R. Mason, and Timothy Streeter

*Department of Speech, Language and Hearing Sciences and Hearing Research Center, Boston University,
635 Commonwealth Avenue, Boston, Massachusetts 02215*

Eric R. Thompson

Ball Aerospace and Technology Corp., Fairborn, Ohio 45324

Virginia Best

National Acoustic Laboratories, Macquarie University, New South Wales 2109, Australia.

Gregory H. Wakefield

*Department of Electrical Engineering and Computer Science, University of Michigan,
Ann Arbor, Michigan 48109*

(Received 16 August 2012; revised 30 May 2013; accepted 8 June 2013)

This study examined the ability of human listeners to detect the presence and judge the strength of a statistical dependency among the elements comprising sequences of sounds. The statistical dependency was imposed by specifying transition matrices that determined the likelihood of occurrence of the sound elements. Markov chains were constructed from these transition matrices having states that were pure tones/noise bursts that varied along the stimulus dimensions of frequency and/or interaural time difference. Listeners reliably detected the presence of a statistical dependency in sequences of sounds varying along these stimulus dimensions. Furthermore, listeners were able to discriminate the relative strength of the dependency in pairs of successive sound sequences. Random variation along an irrelevant stimulus dimension had small but significant adverse effects on performance. A much greater decrement in performance was found when the sound sequences were concurrent. Likelihood ratios were computed based on the transition matrices to specify Ideal Observer performance for the experimental conditions. Preliminary modeling efforts were made based on degradations of Ideal Observer performance intended to represent human observer limitations. This experimental approach appears to be useful for examining auditory “stream” formation and maintenance over time based on the predictability of the constituent sound elements. © 2013 Acoustical Society of America.

[<http://dx.doi.org/10.1121/1.4812276>]

PACS number(s): 43.66.Mk, 43.66.Lj, 43.66.Ba [LD]

Pages: 1215–1231

I. INTRODUCTION

When an acoustic environment is filled with many sources of sound, a listener is faced with a series of challenging problems: First, the various sound sources must be parsed and segregated so that each may be evaluated for relevance. Next, attention must be directed to the source of greatest interest so that whatever information is being conveyed by that source is given the appropriate processing. Over time, attention must be maintained on the highest priority source (noting that “priority” is determined by the observer) while still allowing the remainder of the sound field to be monitored and attention to be redirected at will. Each of these tasks usually must be accomplished under the pressure of competition from unwanted sound sources. Consideration of this latter point reveals an inherent dilemma for the listener who must keep attention focused on one source, ignore

others, yet still monitor “unattended” sources in case the priorities change and an unwanted sound source (the “masker”) becomes the desired sound source (the “target”) with the concomitant need to refocus attention.

Although it certainly is true that there are single, brief acoustic events that convey important information in many situations (e.g., see Yost *et al.*, 1989, for a review of the processing of “transient” sounds) it also is true that many important auditory tasks—such as spoken communication—require maintaining the focus of attention on a specific sound source over extended periods of time linking together *sequences* of acoustic events. The term “auditory stream” has been used to denote a variety of conditions under which sequences of sounds¹ appear to originate from the same source creating in the listener the sense of perceptual coherence (e.g., van Noorden, 1975; review by Bregman, 1990; see also Yost, 1991). The perception of coherence in a sequence of sounds—which broadly means that the sounds appear to be related²—facilitates the focus of attention on a target source and aids in maintaining the integrity of the stream in the presence of competing maskers (e.g., Kidd *et al.*, 1995, 2003; Huang and Richards, 2008; Shamma *et al.*, 2010). The study of stream formation, segregation,

^{a)}Portions of this work were presented at the 35th Annual Meeting of the Association for Research in Otolaryngology, San Diego, CA, February, 2012.

^{b)}Author to whom correspondence should be addressed. Electronic mail: gkidd@bu.edu

and maintenance over time has become integral to the understanding of multisource listening in general, and much attention has been given to how listeners accomplish these feats and, as importantly, why they sometimes fail.

There are different perspectives on what constitutes an auditory “stream.” In Bregman’s (1990) treatise on “auditory scene analysis” he states that “an auditory stream is the perceptual grouping of the parts of the neural spectrogram that go together...” (p. 9) and “I view a stream as a computational stage on the way to a full description of an auditory event...” (p. 10). More recently, Bregman (2008) has defined an auditory stream as “A perceptual interpretation, produced by auditory scene analysis, of a mixed auditory input... (it) is a perceptual entity, analogous to a visual object, capable of binding a cluster of properties...” (p. 861). Shamma *et al.* (2010) define streaming as “The process of forming segregated percepts of auditory sources...” and that “in the hearing-research literature... (streaming) describe(s) sequential grouping or organization of sound segments or tokens over time” (Glossary of terms, p. 10). They also emphasize the important role of temporal coherence of across-frequency sound elements, suggesting that perceptual coherence (due to the repetition of synchronous sound elements) has a strong role in stream formation. Winkler *et al.* (2012) propose that there are two principles that govern how sequences of sounds are bound together to form streams. The first principle is perceptual similarity between sequential events. “Similarity,” like grouping and coherence, is subject to various interpretations but in this application reflects the extent to which successive sounds share one or more common perceptual features or, as proposed by Winkler *et al.*, is related to the rate of change of a feature between successive events. The continuity of, or plausible change in, a simple acoustic feature means that it is likely to reflect the physical properties of a unique source (e.g., material composition, size and shape, etc.; see Lutfi, 2008, for a review). The second principle proposed by Winkler *et al.* (2012) is predictability. Sounds are bound together over time if the listener detects “predictive relations” among them. For example, if a sequence of sounds repeats periodically then the constituent elements are likely to have been generated by a common mechanism. The predictability of the repeating elements causes the observer to experience them as comprising a whole; i.e., they form a coherent structure or stream. Not all conditions that lead to streaming are influenced by variations in predictability, however, as shown for the “fission boundary” in A-B-A tone sequence perception (e.g., French-St. George and Bregman, 1989).

It is perhaps most generally agreed that a sequence of sounds which are related by some consistent feature/physical property, such as the fundamental frequency of voiced speech or a spectral prominence associated with the resonance of a cavity, or a constant spatial location, may form a stream if the listener perceives that feature as a means of binding the sounds together over time. The consistency of this feature may connect sounds even if they vary in other unrelated ways. Again using speech as an example, a plausible pattern of fundamental frequency variation—a feature—can bind speech sounds together even though the individual

phonemes vary in other ways, such as in the dynamic change in spectral shape associated with a formant transition or a plosive burst. However, a compelling argument may also be made that streams can be formed by more complex connections among the sound elements, such as when sequences of speech sounds follow learned syntactic rules. Thus, talkers utter “streams of speech” (e.g., Licklider, 1952) that are bound together in multiple ways including both low-level acoustic features and high-level linguistic structure. It is trivial to demonstrate, for example, that a listener can understand a series of individual words spoken by different talkers—and thus predominantly have disjoint talker-specific features—if the words form a syntactically correct sentence, even under masked conditions such as multitalker competition. The words are bound together by conforming to a learned syntactic structure and by the semantic continuity of the message that is conveyed (e.g., Brouwer *et al.*, 2012), which may be communicated independent of the pattern of variation of low-level features. This observation also seems generally consistent with the idea of higher-level binding mechanisms (similar to what Bregman termed “schema-based grouping”). Conforming to an expected pattern provides a type of “coherence” just as does the plausible variation of a low-level feature (e.g., Kidd *et al.*, 2002). More broadly, then, a stream could be considered as a sequence of auditory events that conforms to an expected relation or learned pattern—which may include the presence of a constant lower-level feature—but may also include more complex connections among sequences of sounds.

Most of the literature on auditory streaming has focused on two extremes: The perceptual coherence of low-level features in simple stimuli, such as in the A-B-A tone sequence demonstration by van Noorden (1975; see also reviews in Bregman, 1990; Darwin and Carlyon, 1995; Micheyl *et al.*, 2007; Shamma *et al.*, 2010; Moore and Gockel, 2012), and the much more complex and multidimensional connections between elements comprising intelligible speech. The reason that these are characterized as extreme cases along a hypothetical continuum of auditory stream phenomena is that they serve very different purposes and convey very different levels and types of information. For an A-B-A sequence, all that is communicated to the listener is that a coherent entity, or entities, exists; it merely elicits a perceptual impression that extends over time. In contrast, a stream of speech may convey any of a very large number of potential meanings in addition to eliciting the impression of a separate perceptual entity and is governed by the predictability of word order via syntactic structure. In both cases, and some that fall in between (e.g., polyphonic music, environmental sounds), it is much simpler to manipulate the variation of low-level features, such as fundamental frequency/pitch or location, than to control the predictability of the elements comprising a sequence, per se, that leads to expectation about arriving/impending sounds.

The purpose of the present study is to examine directly the idea that listeners are able to judge the presence and strength of statistical dependencies among sounds presented in sequence. In order to test this hypothesis, a new empirical approach is proposed in which the probability of occurrence

of a given sound depends on the previous sound. This approach is appealing in that it allows the strength of the predictive connection—the degree of dependency—among sounds to be varied in a precise and controlled way. It also allows for certain quantitative analyses to be developed and applied as a means for understanding the limitations on human performance.

A. Markov chains and specification of the Ideal Observer

In order to examine the role of predictability in stream perception, a means must be found for varying the likelihood of the occurrence of one sound following another in a sequence of sounds. Given a finite set of N states (sound elements, in the present application), a Markov chain is a random sequence of events in which the conditional probability of the state of the current event depends only on the state of the immediately preceding event. The random behavior of a Markov chain is determined by a probability density function for the state of the first event in the sequence and a *transition matrix* of the conditional probabilities, where the (j,k) -th entry represents the probability that the current event is in the k -th state given that the preceding event was in the j -th state. Thus, the j -th *row* of the transition matrix represents the probabilities of going from the j -th state to the 1st, 2nd..., k -th... N -th state, while the k -th *column* of the transition matrix represents the probability of arriving at the k -th state from the 1st, 2nd..., j -th... N -th state. By generating sequences of sounds according to specified transition matrices the strength of the dependency among the elements comprising the sequence may be systematically varied. For example, if all of the conditional probabilities in the j -th row are $1/N$, then it is equally likely to go from a state- j event to any of the other states on the following event. In contrast, if one element in the j -th row is 1 (say, in the i -th column) and the rest are zero, then a state- i event always follows a state- j event in the sequence. The appreciation of such Markov sequential dependencies has long formed a basis for models of speech perception and the related application to algorithms implementing automatic speech recognition (e.g., Rabiner, 1989). Furthermore, previous work employing sound sequences generated according to transition matrices has been reported both in psychophysical (e.g., Pollack, 1970) and physiological (e.g., Mill et al., 2011) studies of auditory processing. The relation between expectation and stream formation and maintenance has been discussed both in theoretical and mechanistic terms (e.g., Denham and Winkler, 2006; Overath et al., 2010; Winkler et al., 2012).

In the present example, there are six states (a, b, c, d, e , and f). These states correspond to the values along a stimulus dimension from which the different sound elements are drawn. These states need not be confined to values along a single, simple, acoustic dimension, however, but could comprise such stimuli as uniquely identifiable nonspeech patterns or actual speech sounds (e.g., Kidd et al., 2012). In this study, the stimulus values are either tonal frequencies referred to as “pitch” states or interaural time differences, termed “location” states, acknowledging that these labels

TABLE I. Sample transition matrices showing the conditional probability of occurrence of each of the six possible states (a - f) in the Markov chains used to form stimulus sequences. The entries in the matrices indicate the probability of occurrence of each state on event $k + 1$ given the state that occurred on the preceding event k . The shaded diagonal indicates the probability, p , of a repetition of state, while the probability of a transition in state equals $1-p$. Table I(a) shows the uniform matrix used for non-target stimulus draws (experiment 1) while Table I(b) shows one of the non-uniform matrices used to generate target stimuli (see text).

		Event $k + 1$					
I(a)		a	b	c	d	e	f
Event k	a	1/6	1/6	1/6	1/6	1/6	1/6
	b	1/6	1/6	1/6	1/6	1/6	1/6
	c	1/6	1/6	1/6	1/6	1/6	1/6
	d	1/6	1/6	1/6	1/6	1/6	1/6
	e	1/6	1/6	1/6	1/6	1/6	1/6
	f	1/6	1/6	1/6	1/6	1/6	1/6
		Event $k + 1$					
I(b)		a	b	c	d	e	f
Event k	a	5/6	1/30	1/30	1/30	1/30	1/30
	b	1/30	5/6	1/30	1/30	1/30	1/30
	c	1/30	1/30	5/6	1/30	1/30	1/30
	d	1/30	1/30	1/30	5/6	1/30	1/30
	e	1/30	1/30	1/30	1/30	5/6	1/30
	f	1/30	1/30	1/30	1/30	1/30	5/6

oversimplify the actual percepts. Table I(a) shows one of the limiting cases of the transition matrix in which all the elements are equal ($1/6$) and represents the degenerate case of an *independent* Markov chain in which the state of the current event is statistically independent of the state of the previous event. In contrast, the matrix shown in Table I(b) is not uniform but has identical probabilities (greater than $1/6$) along the main diagonal and identical probabilities (smaller than $1/6$) along the off-diagonals. Thus, for a Markov chain generated using the transition matrix shown in Table I(b), the state of the current event is the most likely state of the preceding event. This structure imposes a quantifiable degree of sequential dependency among the elements comprising the stream which makes it well-suited as a means for manipulating the independent variable in a perceptual experiment. For such a transition matrix, p determines the probability that there is *no change in state* between adjacent events and $1-p$ denotes the probability that there is a *change in state* between adjacent events in the sequence. The value of $1-p$ will be called the *transition probability* throughout the rest of this paper.

Two such sequences—one drawn as a sample from the matrix in Table I(a) and the other from the matrix in Table I(b)—are illustrated schematically in Fig. 1. The probability of the state of the first event was a uniform distribution over the six states. The states are indicated on the ordinate while the samples for events k_{1-12} are shown over time along the abscissa. For the sequence drawn from the uniform matrix in Table I(a), nine transitions occurred, i.e., the states of adjacent events in the sequence differed. In the sequence drawn from Table I(b), only two transitions occurred, as expected,

FIG. 1. A stimulus schematic illustrating examples of non-target (top panel) and target (bottom panel) sequences. The abscissa is time (element or event number) and the ordinate is stimulus state (frequency or ITD). The transition matrices that generated these examples (Table I) were independent/uniform (top) or dependent/non-uniform (bottom).

since the likelihood of staying in the current state (p) is much greater than $1/6$. In the experiments that follow, the listener's task was to determine whether or not the test sequences were drawn from a non-uniform transition matrix (experiment 1) or, in cases where pairs of sequences must be compared (experiments 2 and 3), which sequence was drawn from the matrix having the lower transition probability. The logic behind designating the target in this manner is based on the assumption that streams containing information that is useful or important to a human listener exhibit sequential statistical dependencies; that is, they are not a collection of independent, random events but rather follow a particular conditional dependence, structure, or pattern³ between adjacent events. Note that the stimulus dimension(s) along which the states are established does not matter with respect to the underlying probability distributions associated with the transition matrices. In all cases, too, the differences in transition probabilities inherently affect how numerous the transitions are which provides evidence that may be used in solving the discrimination task. However, the choice of stimulus dimension—and indeed the details of stimulus presentation—has perceptual consequences that may affect human discrimination performance. For example, if the listener were unable to discern differences between the physical states used to define the sound elements, they would not be able to discriminate differences among the transition matrices used to generate Markov chain sequences.

Once the transition matrices from which the stimuli are drawn are established, it is possible to compute the performance of an Ideal Observer. Given an observed sequence of sound elements, the Ideal Observer bases a decision on the likelihood ratio constructed from knowledge of the discrete distributions defined by the transition matrices.⁴ In the experiments that follow, Ideal Observer performance is presented as a standard against which human performance is compared. For the conditions relevant to these experiments, Ideal

Observer performance depends on the probability distribution defining the initial state in the sequence, which as indicated above is uniform across the set of states in all conditions tested here, the transition matrices governing the stimulus draws, and the number of events comprising the test sequence.

The following experiments are intended to examine several aspects of sequential perception that relate to stream formation and maintenance. First, are listeners able to detect and judge the strength of statistical dependencies in sequences of sounds, as represented by the transition probability, $1-p$? Is this a general ability that would be apparent for stimuli distributed along different perceptual dimensions? Second, if a statistical dependency between adjacent events is present for one feature in a stimulus, can listeners relate sound elements connected by that target feature while a second irrelevant feature varies independently between adjacent events? And finally, how well can listeners evaluate multiple concurrent sound streams; that is, can listeners simultaneously examine two sequences of sounds and determine which has a stronger statistical dependence among its adjacent elements? In each case, we also sought to determine and specify Ideal Observer performance for the conditions tested, then later to consider how human performance differed from Ideal and the reasons for those differences.

II. GENERAL METHODS

The test subjects were young adult volunteers with normal hearing as determined by routine audiometric screening. Five subjects participated in experiment 1, six subjects in experiment 2, and eight subjects in experiment 3. One of the subjects in experiment 2 also participated in experiment 3. The subjects were tested individually in double-walled, sound-attenuating IAC booths while seated at a station with a video display monitor, keyboard, and mouse. The stimuli were presented at 60 dB sound pressure level (SPL) through Sennheiser 280 Pro earphones. The subjects were given correct answer feedback after each trial.

All sounds were generated at a sampling rate of 50 kHz, low-pass filtered at 20 kHz and presented through TDT System II hardware (Tucker-Davis Technology, Inc.). In all conditions the sequences of sounds were generated according to transition matrices similar in form to the examples shown in Tables I(a) and I(b) with the first element chosen randomly from the set of states. The matrices were always either the $m \times m$ uniform matrix with all entries equal to $1/m$ (where m is the number of states) or a non-uniform $m \times m$ matrix with the number p on the main diagonal greater than $1/m$. The length of the sequences was 12 elements and there were either 6 (experiments 1 and 2) or 7 (experiment 3) states from which the Markov chains were constructed. The states were either different frequencies (“pitch”, all 3 experiments) or different interaural time differences (ITDs; “location”, experiments 1 and 2).

III. EXPERIMENT ONE: DETECTING THE PRESENCE OF A STATISTICAL DEPENDENCY IN A SEQUENCE OF SOUNDS

The first question is whether listeners are able to detect the presence of a statistical dependency among the elements

comprising a sequence of sounds. To address this question, we employed sequences in which the 12 elements were drawn from the two types of transition matrices described above for the perceptual dimensions of pitch and location. The stimuli drawn from the non-uniform matrices were designated as the “target” sounds, and those from the uniform matrix were the reference or “non-target” sounds.

A four-fold difference in duration was examined to provide an initial determination of the role of time (i.e., presentation rate and element duration) in discriminating sequences of this type. Informal listening suggested to us that there were perceptual differences in the sequences due to duration/rate that might lead to differences in performance. However, the underlying statistical variation of the elements (states) is equivalent for these variables and so too are the predictions of the Ideal Observer.

A. Stimuli

There were two sets of stimuli used in this experiment. One set comprised six pure tones having frequencies ranging from 1000 to 1250 Hz spaced at equal intervals on a logarithmic scale. The second set of stimuli consisted of independent noise bursts drawn from a set of six ITDs (± 100 , ± 500 , and $\pm 1000 \mu\text{s}$). Two durations were tested for each type of stimulus: 60 ms and 240 ms, interleaved with silent gaps of equal duration. Thus, the total duration of the 12-element sequence within a trial was 1.38 s for the briefer stimulus and 5.52 s for the longer stimulus. Each sound in a sequence was gated on and off with 10-ms cosine-squared ramps.

B. Procedures

The task was one-interval forced-choice (i.e., “YES-NO”). The subjects were initially instructed about the task in a general way—stimulus timing, response mode, display, etc.—and were told to learn to solve the task by attending to a series of stimulus examples (described below) and to the trial-by-trial response feedback. No additional information was provided about the nature of the sounds. A response was counted correct if the listener indicated “YES” when the stimulus was drawn from a non-uniform transition matrix (the “target”) or “NO” if the sequence was drawn from the uniform transition matrix. The independent variable was the transition probability of the target and values of 0.7, 0.5, 0.3, and 0.1 were tested. The non-target stimuli had a transition probability of about 0.83 (5/6). Prior to each block of trials, the listeners heard 16 tokens from the target and non-target distributions (eight from each, two of each target transition probability) played as examples of the stimuli that would be used in that block. The stimuli were identified as targets or non-targets with no response required by the listeners. Each test block consisted of 48 trials equally divided between target and non-target presentations. The target trials were equally distributed among the four transition probabilities. Each subject completed 12 blocks for each combination of the two state dimensions and the two durations, for a total of 48 test blocks yielding 72 trials per subject per data point (state dimension, duration and transition probability). These

data were obtained over four sessions of about 2 h each per subject. Each 2-h session consisted of testing one state dimension type (two subjects started with pitch and three subjects started with location). Within each session, each duration was tested in one-half of the blocks.

C. Results

An important issue for the design of this study was how quickly and accurately listeners were able to learn the underlying statistical structure—the distributions of the target and non-target stimuli—based on the exemplars and response feedback that was provided. The performance observed in experiment 1 was examined as a function of block number, and these learning functions (averaged across subjects and transition probability values) are shown in the left panel of Fig. 2.

In general, the subjects rapidly learned the task during the first block(s) of trials with very little improvement apparent over subsequent blocks. Averaged across all four conditions, the group-mean value for the first three blocks of trials was 69% correct, while the mean value for the final three blocks of trials was 73% correct with the difference due largely to the improvement during the first block of trials. Greater improvement was seen for the brief duration stimuli, which increased about ten percentage points over the course of testing (again, primarily during the first block) while virtually no improvement was seen for the longer duration stimuli. A repeated-measures analysis of variance (ANOVA) was conducted on these data to determine whether performance improved significantly over this range of practice. Because overall performance for pitch and location states is related to the somewhat arbitrary state values chosen for each, separate ANOVAs were performed. For the pitch

FIG. 2. Group mean results from experiment 1. The ordinate is percent correct performance. In the left panel, the abscissa is the trial block number while in the right panel the abscissa is transition probability. The four functions with data points indicate the two stimulus dimensions, pitch (frequency; triangles) and location (ITD; circles). Filled symbols with solid lines indicate brief durations and open symbols with dashed lines indicate long durations. The error bars indicate ± 1 standard error of the mean. The dashed-dotted line at the top of the right panel shows Ideal Observer performance while chance is shown as the dotted line at 50% correct.

variable, the main factor of duration was significant [$F(1,4) = 11.25$, $p = 0.028$] whereas block number was not [$F(11,44) = 1.69$, $p = 0.107$]. The interaction of duration and block number was also significant [$F(11,44) = 2.49$, $p = 0.029$]. This interaction reflects the relatively large improvement found for the short-duration tones initially. For the location data, duration was not significant [$F(1,4) = 1.05$, $p = 0.36$] while block number was significant [$F(11,44) = 2.24$, $p = 0.029$]. The interaction also was not significant [$F(11,44) = 1.66$, $p = 0.11$]. A general observation is that the procedure of training the subjects through the explicit instructions to monitor response feedback, coupled with the examples given of tokens from target and non-target distributions, appears to have facilitated training for these conditions. This suggests that listeners are very sensitive to statistical dependencies among adjacent events in sound sequences and that this does not require extensive training. However, it should be noted that learning more complex statistical dependencies might require more practice and exhibit greater improvement over time than was found here.

The right panel of Fig. 2 contains group-mean percent correct scores and standard errors of the means plotted as a function of transition probability for both durations. Because no pronounced improvement in performance was found after the first block of trials, all subsequent blocks were used in tabulating average scores. Performance was better than chance for all conditions tested and increased monotonically with decreases in transition probability for both pitch and location conditions. In general, performance was better when the set of elements differed along the dimension of pitch (frequency; triangles) than along the dimension of location (ITD; circles). Furthermore, performance was slightly better along both dimensions for the longer element/sequence duration (open symbols). The psychometric functions were slightly steeper for pitch than location as well over this range of transition probabilities. However, these comparisons are made with the caveat that performance may very well depend on the specific values that were chosen to create the sets of states and that different values might have led to different levels of performance. In the current experiment, we did not attempt to match the intervals separating the values in frequency and ITD according to some uniform criterion such as number of just-noticeable-differences (jnds).

A repeated-measures analysis of variance was conducted for pitch and location variables separately, as above. For the pitch dimension, transition probability [$F(3,12) = 262.9$, $p < 0.001$] was significant while duration [$F(1,4) = 1.7$, $p = 0.26$] was not. The interaction of transition probability and duration was also significant [$F(3,12) = 4.07$, $p = 0.033$]. For the location dimension, the main effect of transition probability was significant [$F(3,12) = 44.5$, $p < 0.001$] whereas duration was not [$F(1,4) = 0.39$, $p = 0.56$]. The interaction of transition probability and duration also was not significant [$F(3,12) = 0.09$, $p = 0.96$].

Note that chance performance for the Ideal Observer occurs at a transition probability of 0.83 (not shown), which reflects the case when target and non-target matrices are the same (uniform matrix of 1/6; transition probability = 5/6).

For the conditions tested in this experiment, observed human performance was substantially poorer than that calculated for the Ideal Observer.⁵

IV. EXPERIMENT TWO: RANDOM VARIATION ALONG AN IRRELEVANT STIMULUS DIMENSION

The findings of experiment 1 supported the proposition that listeners can make reliable judgments about the presence of a statistical dependency between adjacent events in sequences of sounds distributed along a simple stimulus dimension. Moreover, the accuracy of these judgments reflects the strength of that dependency in an orderly way. This led to the hypothesis that listeners should be able to discriminate the strength of the statistical dependency by directly comparing two or more samples of sequences generated from different transition matrices. In order to test this idea, the next experiment examined performance when the listener was required to make within-trial judgments about the strength of the statistical dependency underlying two successive sound sequences. Furthermore, as noted in Sec. I, complex sounds often vary simultaneously along multiple acoustic dimensions, only a subset of which comprise the features that reliably bind sequential elements together over time. Thus, one property may help connect sound elements into streams while other properties do not and may potentially confound the stream formation/maintenance process. To explore this possibility empirically, conditions were tested that were intended to examine how well listeners could focus on one particular aspect of a sound sequence, distributed along a target stimulus dimension, while ignoring variation along an irrelevant dimension. In this experiment, the ability of listeners to judge the strength of the target statistical dependency (lower transition probability) along a primary stimulus dimension was compared when the stimulus was fixed within a sequence along a secondary stimulus dimension or was statistically independent between adjacent events.

A. Stimuli

Each trial had two observation intervals of equal duration. Each interval contained a stimulus sequence consisting of 12 contiguous tone bursts (i.e., no silent interval between bursts). Each tone burst was 120-ms duration and the stimulus duration was 1.44 s. The stimulus set consisted of tones drawn from a set of six frequencies ranging from 500 to 630 Hz and a set of six ITDs: ± 100 , ± 400 , and $\pm 800 \mu\text{s}$. Thus, each element (tone) was represented along two perceptual dimensions: Pitch and location. Because these dimensions could be fixed or randomized independently within a sequence, it was possible to vary either or both as controlled variables in the discrimination experiment.

B. Procedures

The task was two-interval two-alternative forced-choice (2I-2AFC) with response feedback given after every trial. The subjects were instructed to use the feedback to learn the property determining the “target” interval (lower transition

probability) in each trial. The two sequences were drawn from different transition matrices. The transition probabilities were 1/6, 2/6, 3/6, 4/6, and 5/6 which yielded ten pairs and four unique paired differences (1/6 through 4/6). Thus, for example, there were four pairs of transition probabilities that had a difference of 1/6 while only one pair had a difference of 4/6; however, each paired difference was presented the same number of times.

Five conditions were tested with each condition fixed throughout a block of trials. One stimulus dimension—frequency or ITD—was designated as the target dimension—while the other dimension was designated as the nontarget/secondary dimension, except for the correlated condition described below in which both dimensions were target dimensions. The subjects were informed about the test condition prior to each block. If the secondary dimension was fixed, then on each interval of every trial one of the set of six states was selected at random and each tone in the sequence was presented at that value. If the secondary dimension was variable, each of the nontarget dimension states was selected independently from the set of six for each tone burst in a sequence, for every sequence in each interval of every trial. In the correlated case, frequency and ITD covaried such that high-to-low frequency was perfectly correlated with left-to-right leading ITD. In that case, either pitch or location—or a combination of both—could be used to solve the task. The different conditions are represented schematically in Fig. 3.

Each row of panels illustrates the variation of the target sequence along the target (left panel) and secondary/nontarget (right panel) dimensions. Subjects each completed 9 blocks per condition for a total of 45 blocks spanning several

different visits and taking a total of about 4.5 to 5.5 h per subject. The five conditions occurred once in every five blocks in mixed order. Within each block, there were 48 trials with each of the four probability differences presented 12 times. The first block of each condition was considered as practice and was discarded. The data that are plotted and analyzed statistically are thus based on 96 trials per subject per condition per probability difference.

C. Results

The results of experiment 2 are shown in Fig. 4. Group mean performance is plotted in percent correct as a function of the difference in transition probability between the sequences in the two observation intervals within a trial. In all cases, performance was reliably better than chance (dotted line at 50% correct). The lines connecting symbols represent human listener performance while the dashed-dotted line at the top shows performance of the Ideal Observer. For all five conditions tested, listener performance increased monotonically with increasing difference in transition probability. For the target dimension of location performance ranged from about 57–61% correct at the smallest transition probability difference to about 68–73% correct for the greatest difference. Pitch performance was better in all cases, ranging from about 65–67% correct at the transition probability difference of 1/6 to 92–96% correct at the transition probability difference of 4/6. The correlated condition results fell very near to the pitch-alone discrimination functions at each transition probability difference.

A repeated-measures analysis of variance revealed significant main effects of transition probability difference [$F(3, 15) = 59.7$; $p < 0.001$], target dimension [$F(1, 5) = 70.2$;

FIG. 3. Schematic in the same form as Fig. 1 illustrating variation of the elements comprising the test sequences along target (task-relevant “primary”; left column) and non-target (task irrelevant “secondary”; right column) dimensions. The top row illustrates a constant value for the secondary dimension throughout a sequence, the middle panel illustrates random variation of values on the secondary dimension during the sequence, and the lower panel illustrates the case where the two variables covary and thus both are “primary” (target) dimensions.

FIG. 4. Group mean results from experiment 2. The abscissa is the difference in transition probability between the matrices from which the target and non-target sequences were drawn. The ordinate is percent correct discrimination. The error bars are standard errors of the means. The five functions with data points are group means from the different conditions identified in the key (target dimension/secondary dimension). Ideal Observer performance is plotted as dashed-dotted lines with no data points while chance performance is shown as the horizontal dotted line at 50% correct. The filled symbols indicate performance when the secondary dimension was fixed while the open symbols indicate performance when the secondary dimension varied randomly. The stars show performance when target and secondary dimensions covaried.

$p < 0.001$], and randomization of secondary dimension [$F(1,5) = 10.3$; $p = 0.024$]. The only interaction reaching significance was between target dimension and transition probability [$F(3,15) = 12.5$; $p < 0.001$].

The difference in performance due to target dimension was quite large for these conditions. At the higher transition probability differences, where performance was best, the listeners' scores were some 20–25 percentage points higher for a difference in pitch than in location. This is a much greater difference than was found in experiment 1, where performance for transitions in the pitch of tones was compared to transitions in location of noise bursts. Although the jnd for ITD tends to be slightly smaller for noise bursts than for tone bursts (e.g., Klumpp and Eady, 1956; review in Durlach and Colburn, 1978), the jnds are much smaller than the values used here for representing location states.

The effect of randomizing values along the secondary dimension was relatively small but statistically significant. Differences that averaged about 3 percentage points were observed for both location and pitch judgments. Furthermore, correlated variation across the two dimensions (Covary) was not significantly different than the best case performance for pitch alone with constant location. Chance performance for

the Ideal Observer occurs at a transition probability difference of 0 (not plotted). The performance of the Ideal Observer equals that which would be expected from a decision rule in which the sequence having fewer transitions within a given trial is chosen, with random guessing if the number of observed transitions was equal for the two sequences. Human performance fell short of Ideal Observer performance for the values tested although, perhaps because of the ceiling on performance (100% correct) at the upper end, human and Ideal Observer performance was relatively close for the larger transition probability differences along the pitch dimension.

Although the smaller values of transition probability difference occur for pairs of matrices having different absolute values (e.g., 1/6 vs 2/6, and 4/6 vs 5/6) a trend for slightly better performance was observed (in a *post hoc* analysis not shown) for the lower absolute probability differences; i.e., discriminating 1/6 vs 2/6 was better than 4/6 vs 5/6. The same trend is apparent in a breakdown of the Ideal Observer analysis. However, once the (human) data were subdivided in this manner the variability was large enough that strong conclusions were not warranted.

V. EXPERIMENT THREE: SUCCESSIVE VERSUS CONCURRENT COMPARISON OF STREAMS

In typical listening situations, the listener must segregate multiple sound sources and evaluate the streams of sound emanating from each in order to determine where attention should be directed. However, while accomplishing this task, the listener also must remain vigilant with respect to the composition of the sound field so that important changes (e.g., new sources) are recognized and processed rapidly and appropriately. Although there have been studies examining the ability of listeners to perform two auditory tasks simultaneously under conditions of high uncertainty (e.g., Best *et al.*, 2006; Gallun *et al.*, 2007; Ihlefeld and Shinn-Cunningham, 2008; Helfer *et al.*, 2010), comparatively little is known about how listeners perceive and evaluate two concurrent streams of sound.

In the preceding experiments the sound streams were presented in isolation without consideration of the potentially detrimental effect of masking from unwanted sounds or the need to divide processing resources so as to monitor multiple simultaneous streams. In experiment 2, we examined how well listeners could compare two successive streams and judge the degree of sequential dependency among the sound elements. In this experiment, the purpose was to determine the costs of attempting to evaluate the two streams simultaneously. To avoid energetic masking or problems in the perceptual segregation of the two Markov chains into separate streams, the two sequences were presented to different ears and in different frequency regions. The goal was to determine the cost of simultaneous comparison of well-segregated streams compared to comparison of non-simultaneous streams.

Because the sound streams that must be compared in this experiment were separated both by frequency region and by ear, the observer could use either distinction as a means

for indicating their response. That is, if the target stream (as before, the sequence with the lower transition probability) were presented to the left ear and fell in the low-frequency region (with the nontarget stream located in the complementary ear/frequency band), the response could be reported either as “left ear” or “low frequency.” There is little reason to expect that one type of report or the other would be superior since either distinction alone is intended to be sufficient to separate the streams. However, it is possible that report type *uncertainty* might adversely affect performance. Lack of prior knowledge of the channel—ear or frequency band—that must be reported as containing the target could affect the ability to compare the stimuli, especially, we speculated, under concurrent presentation where the process of separating sounds and attaching labels to them would occur during the time that the stimuli must be observed. In order to test this idea, the type of report—ear or frequency band—was designated during each trial either before or after the stimulus presentation.

A. Stimuli

As in experiment 2, there were two sequences of 12 contiguous pure tones presented on every trial and the duration of each tone was 120 ms for a total sequence duration of 1.44 s. Under one condition, the sequences were presented in two successive intervals separated by 500 ms and in the other the sequences were presented concurrently in one interval. The tones comprising each sequence were drawn from two sets of seven frequencies falling within relatively low (800–944 Hz) or high (1500–1770 Hz) frequency ranges. On every trial, one sequence was drawn from the low-frequency band and the other sequence from the high-frequency band. The low and high frequency sequences were then assigned randomly one to each ear. When the sequences were played concurrently, the tone bursts in each were synchronous.

B. Procedures

The task was two-alternative forced-choice. The subjects were instructed to use the response feedback to learn to solve the task. All of the trials in a given block were either successive or concurrent. In one-half of the blocks the listener was informed before each trial which type of response to give (ear or frequency band) and in the other half they were informed after stimulus presentation. Prior to each experimental session, a short response type familiarization procedure was conducted that consisted of four blocks of ten trials each, in which the listeners were presented with only a single sequence that was either the high-frequency band or the low-frequency band in either the right ear or the left ear. They were asked to indicate the ear or the frequency band of the sequence without judging any other aspect of the sounds (i.e., degree of statistical dependency). In two of these training blocks they were informed about response type—and thus which feature to attend to—before the trial and in the other two blocks they were informed about response type after the trial. Response feedback was provided and all subjects quickly demonstrated a high degree of accuracy in this simple task.

As in experiment 2, the two sequences in each trial were composed according to different transition matrices, with the independent variable being the difference in transition probability. Also as in experiment 2, the same five transition probabilities yielding four possible differences: 1/6, 2/6, 3/6, and 4/6, and thus ten pairs of probability differences were tested. The blocks contained equal numbers of trials for each of the four probability differences and for each response type, ear or frequency band. Each subject completed five sessions (on different days) consisting of (1) the 40 trial ear/frequency band identification task, (2) eight blocks of 48 trials each for the successive comparison, and (3) eight blocks of 48 trials each of the concurrent comparison. Each session took approximately 2 h to complete. Therefore, over the five sessions, a total of 20 blocks for each cue condition (pre or post) were obtained for each presentation mode (successive or concurrent) yielding 120 trials per subject per probability difference per response type (ear or frequency band). The data that are presented are based on the last 96 trials after eliminating the first 24 trials in every condition as practice.

C. Results

Figure 5 shows group mean percent correct scores and standard errors of the means as a function of the difference in transition probability. The left panel contains the results for the successive presentation condition while the right panel shows the results from the concurrent presentation condition. Ideal Observer performance and chance performance lines are also indicated by the dashed-dotted and dotted lines, respectively. First, for the successive presentation

FIG. 5. Group mean results from experiment 3. The abscissa is the difference in transition probability between the matrices from which the target and non-target sequences were drawn. The ordinate is percent correct discrimination. The error bars are standard errors of the means. The different conditions are indicated by the symbol key. The type of report—by ear or by frequency band—is indicated together with whether the type of report was designated before (pre) or after (post) the stimuli. The left panel shows performance when the two test sequences were presented successively while the right panel shows performance when the two test sequences were presented concurrently. Ideal Observer performance is indicated as dashed-dotted lines and chance performance is the horizontal dotted line at 50% correct.

condition, performance varied from around 65–70% correct for a transition probability difference of 1/6 to about 90–94% correct for 4/6 difference. The corresponding percent correct scores for concurrent presentation ranged from about 55–64% at the smallest transition probability difference to between about 72–87% at the largest difference. The different report types: Frequency band vs ear, had no systematic effect on performance for successive presentation. For concurrent presentation, reporting target frequency band generally was poorer than the corresponding values for reporting ear of presentation (65.2% vs 72%, respectively, averaged across all cue condition and probability difference values). Pre- vs post-stimulus prompting for report type had no systematic effect under either successive or concurrent presentation conditions.

A repeated-measures analysis of variance on the successive presentation condition revealed that probability difference [$F(3,9) = 108.1$; $p < 0.001$] was significant whereas neither report type [$F(1,3) = 0.48$; $p = 0.54$] nor pre- vs post-stimulus cue [$F(1,3) = 0.048$; $p = 0.84$] was significant. None of the two-way interactions were significant. A repeated-measures analysis of variance on the concurrent presentation condition revealed that both probability difference [$F(3,9) = 33$; $p < 0.001$] and report type [$F(1,3) = 17.5$; $p = 0.025$] were significant whereas pre- vs post-stimulus cue [$F(1,3) = 7.8$; $p = 0.068$] was not. The two-way interaction of probability difference and report type was also significant [$F(3,9) = 7.3$; $p = 0.009$]. Thus, the effect of report type depended on whether the sequences were successive or concurrent with poorer performance found in the concurrent case when the report was according to frequency band. The hypothesis that pre- vs post-stimulus cue mattered for concurrent presentation was not supported.

The poorer performance for concurrent as compared to successive presentation (the “cost” of simultaneous comparison) is plotted in Fig. 6 as the group mean difference (successive minus concurrent) in percent correct performance averaged across transition probability differences. The costs of concurrent presentation tended to be larger when the report was by frequency band (about 15 percentage points) than by ear (about 9.5 percentage points) with the smallest difference occurring when the target feature to be reported was designated prior to the stimulus. The highest costs of concurrent presentation overall occurred when frequency band was reported for the larger transition probability differences. In that case, the difference in group mean performance (comparing across panels in Fig. 5) averaged 18 percentage points for the differences of 3/6 and 4/6.

Although the two sound sequences presented concurrently in the current experiment were separated both by frequency and by ear, the argument could be made that the repeated synchronous gating of the elements of the two sequences promoted temporal coherence, fostering some degree of perceptual grouping. That is, the poorer performance in concurrent presentation than in successive presentation might be due to a loss of detail in the processing of each sequence because they were grouped into a single stream. As noted in Sec. I, Shamma *et al.* (2010) have suggested that the repetition of synchronously presented sound elements

FIG. 6. The difference in group mean percent correct performance from experiment 3 averaged across all transition probability differences. The bars indicate the magnitude of the difference averaged across subjects with error bars indicating standard errors of the means. The left pair of bars shows the differences when the report mode was indicated before the stimulus (pre) while the right pair of bars shows the differences when the report mode was indicated after the stimulus (post). Filled bars indicate the differences when the report mode was frequency band while the open bars are the differences when the report mode was ear. A positive difference means better performance in the successive condition.

(e.g., tones in different frequency regions) fosters the binding of the elements into a common stream. Recently, Micheyl *et al.* (2013) considered a “strong interpretation” of grouping due to temporal coherence, arguing that such an interpretation would mean that repeating across-frequency sound elements—causing perceptual coherence—would result in “obligatory” streaming. Extracting the properties of the two sound sequences when perceptually grouped together would be more difficult—perhaps even impossible—than when they were separated in time. Thus, the decrement in performance found under the conditions just described could be an indication of the loss of stimulus detail due to obligatory grouping rather than a reflection of the cost of concurrent evaluation of streams, *per se*.

In order to evaluate this hypothesis, an additional condition was tested in which the tonal elements in the two sequences in the concurrent presentation case did not have synchronous onsets/offsets. This experiment was identical to the concurrent condition described in the preceding section with the following exceptions: (1) A new group of 4 normal-hearing young-adult subjects was tested; these subjects had not participated in any of the other experiments in this study; (2) only the pre-cue condition was tested; (3) a total of 96 trials per point were obtained over the 2-h sessions with the first 24 trials discarded as practice; as a consequence of these design differences statistical comparisons across the two groups (previous conditions described above and current conditions) were not attempted; and (4) one of the two sequences (chosen at random) was delayed by 1/2 the duration of each element, i.e., by 60 ms. Thus, a large asynchrony

was present between the sequences. If the difference between successive and concurrent presentation were due to obligatory grouping, as postulated in the preceding section, then we would expect that asynchronous presentation would further segregate the two sequences leading to substantially better discrimination performance in the concurrent presentation.

The results of this new condition, along with portions of the data shown in Fig. 5 replotted for comparison, are contained in Fig. 7. For both the previous results and the current results (these are designated “asynchronous” to distinguish them from the previous results even though that designation was not relevant for successive presentation), the dimension for report (either frequency band or ear) was identified before the trial and so only those data are shown in Fig. 7. Note that the two panels of Fig. 7 are for different modes of report, frequency band or ear. Ideal Observer performance is the same in both panels and is the same as shown in Fig. 5. The dark solid lines and symbols with error bars indicate the results (intersubject means and standard errors)—including asynchronous-concurrent presentation and the replication of successive presentation—while the light dashed lines and symbols (error bars shown in Fig. 5 but not here) show the subset of data replotted from Fig. 5. The left panel shows the results for report by frequency band. Here performance was generally poorer for concurrent presentation and some differences between asynchronous and synchronous conditions were found. For successive presentation, group mean performance averaged across all transition probability

differences was 80% correct for the new subject group while it was about 82% correct for the previous subject group. For the asynchronous-concurrent condition reported by frequency band, the average performance was about 72% correct whereas it was about 68% correct for the synchronous-concurrent condition. The right panel shows the results for report by ear. In both successive and concurrent presentation, the patterns of results were very similar to those obtained for the previous conditions. Averaged across all subjects and transition probability differences, the mean percent correct for the successive presentation was 84% while it was 82% previously. For the asynchronous-concurrent presentation, the average percent correct was about 77% while for synchronous-concurrent it was about 75%.

A repeated-measures analysis of variance was conducted on the successive and asynchronous-concurrent results. Significant main effects were found for probability difference [$F(3,9)=41.6$; $p<0.001$], report type [$F(1,3)=74.6$; $p=0.003$] and successive vs asynchronous-concurrent presentation [$F(1,3)=13.6$; $p=0.035$]. The 2-way interactions of probability difference and report type [$F(3,9)=4.42$; $p=0.036$] and probability difference and successive vs concurrent presentation [$F(3,9)=0.021$] also were significant.

The important point for the asynchronous-concurrent presentation condition was that performance remained significantly poorer than for successive presentation. This supports the earlier finding for synchronous-concurrent presentation described above. However, the magnitude of the decrease in performance relative to successive presentation was less for asynchronous than for synchronous presentation. The possible reasons for this difference are considered in the next section.

VI. DISCUSSION

A. Summary of experimental findings

Based on the results described above, it is clear that human listeners can reliably detect the presence of a statistical dependency between adjacent events comprising a sequence of sounds. This conclusion is based on the evidence obtained in the first experiment where the dependency was imposed on the perceptual dimensions of pitch or apparent location of the sounds. The observed human performance for both corresponding physical variables, frequency and ITD, yielded orderly psychometric functions that increased monotonically with decreasing transition probability over the range tested. The similarity of the findings for location of noise bursts and pitch of tones, which produce very different percepts in the listener, is notable because it implies that the discriminability of statistical dependencies between adjacent events in a sequence is not limited to a specific stimulus or its perceptual effect. For the types of stimuli and values employed in the experiment, which were chosen to be clearly discriminable but not equated otherwise, performance for the perceptual dimension of pitch was superior to that found for the dimension of apparent location. Listeners learned to make these discriminations rapidly and accurately. In addition, performance in this YES-NO

FIG. 7. Group mean results from the second part of experiment 3 comparing performance under successive and asynchronous-concurrent presentation (labeled asynchronous-successive and asynchronous-concurrent). The abscissa is the difference in transition probability between the matrices from which the target and non-target sequences were drawn. The ordinate is percent correct discrimination. The results are plotted by bold symbols connected by solid lines. The error bars are standard errors of the means. The two panels contain data for the different types of report, frequency band or ear. For comparison, the light gray symbols connected by dashed lines show the data, for report mode given pre-trial, from the first part of experiment 3 (replotted from Fig. 5; see text). Ideal Observer performance is indicated as heavy dashed-dotted lines and chance performance is the horizontal dotted line at 50% correct.

detection task did not vary significantly for a four-fold change in stimulus duration. This stimulus parameter was examined due to the intuition that different detection strategies might be used by the observer for different rates of presentation of the sound elements. Although the negative results were not conclusive since they are limited to the specific range of values tested, no support was found for the underlying hypothesis.

The second experiment found that listeners were able to discriminate between two successive sound sequences that differed in the strength of the dependency among elements.

To accomplish this task, listeners presumably made some judgment about the first sequence, held that judgment in memory, and then compared it to the subsequent stimulus. The underlying basis for this judgment is not readily apparent from these data. For this 2I-2AFC task, Ideal Observer performance is equivalent to choosing the sequence having fewer observed changes of state regardless of the stimulus dimension. However, performance differed for pitch and location, and both were less than that of the Ideal Observer, which suggests that stimulus dimension-specific factors may contribute to the limitations on performance.

For this discrimination task, random variation along a secondary—task irrelevant—stimulus dimension produced small but statistically significant decrements in performance relative to fixed values of the stimulus along the secondary dimension. This may be interpreted as indicating that the listener was unable to exclusively attend to the target dimension. It is even possible that listeners are obliged to consider both dimensions despite instructions to disregard one of them. The coupling of relevant and irrelevant features in a speech identification task was examined recently by [Maddox and Shinn-Cunningham \(2012\)](#) using spoken digits that varied in fundamental frequency and apparent location. They presented pairs of digits in sequence that had the same or different values of pitch or location and subjects were required to use one or the other to select the target digit. The other non-target feature was irrelevant—even misleading—for solving the task and the listener was instructed to ignore that feature. Their results suggested that relevant and irrelevant features tend to be bound together and, the greater the separation between sequential elements along the task-irrelevant dimension, especially pitch, the more interference it produced. Despite the differences in stimuli and task, our results appear to be broadly consistent with those of [Maddox and Shinn-Cunningham \(2012\)](#). Although the detrimental effect we found here was significant, it was surprisingly small. In particular, the ability to discriminate the sequential dependency of pure tones in location even when pitch varied randomly was unexpected given the large effect variations in pitch can have on other detection and discrimination tasks (e.g., [Watson, 2005](#)). However, it should be noted that the difference in variability along the secondary dimension between the fixed and random conditions tested here was confined to variation within vs across stimulus intervals. Even in the fixed condition, the value of the secondary variable changed unpredictably from sequence to sequence and trial to trial—it was only “fixed” within a sequence. Thus, it is possible that a greater influence on performance would

have been found if the secondary variable had been even more stable, e.g., was a constant value across intervals within and across trials in the fixed case.

The superiority in discriminating sequential dependencies in pitch compared to location was more pronounced in experiment 2 than in experiment 1. This result may be attributed to differences in the design of the experiments or, more specifically, to differences in the stimuli that were used. Experiment 1 employed a one-interval forced-choice detection task while experiment 2 used a two-interval forced-choice discrimination task. In experiment 1 the non-target transition matrix was uniform while the target transition matrix was not. Thus, in experiment 1, the judgment made by the listener presumably was based on knowledge acquired about the underlying distributions of target and non-target stimuli. A decision then would be based on where the stimulus fell along the decision axis relative to an established criterion held in a memory store that persisted at least across the block of trials. In contrast, presenting two stimulus samples on consecutive intervals within a trial allowed direct comparison of the stimuli placing somewhat different demands on memory (e.g., [Harris, 1952](#); [Braidia and Durlach, 1972](#); [Berliner and Durlach, 1973](#); [Kidd *et al.*, 1988](#)). The different procedures yielded roughly comparable performance for judging dependencies along the pitch dimension for the two experiments, but not for location, so we cannot rule out an interaction between the task and the perceived stimulus dimension. Also, the subjects were different in the two experiments, and simple intersubject differences in the ability to discriminate location may be a factor. The set of ITD values tested spanned a slightly smaller range too which may contribute to the poorer performance in experiment 2. Finally, however, the most likely cause would seem to be the different stimuli that were used to form the location states—tones vs noise. Even though the tones were fixed in frequency within a sequence they still varied from sequence to sequence thus causing significant frequency uncertainty that must be ignored while making location judgments. As noted above, this variability even in the most certain cases could have influenced the judgments along the primary dimension when the stimuli were tones. For noise, though, this would not have been the case and this may be the reason for the better performance found in experiment 1 for location. It is also possible that rapid changes in location—rather than acuity for a single event—are represented in the binaural auditory system less accurately; i.e., greater “binaural sluggishness” (e.g., [Culling and Colburn, 2000](#); [Hall *et al.*, 2011](#)) which could be different for rapid sequences of noise bursts than for tones or for the different subject groups. However, at present the basis for the difference in performance across the two experiments is uncertain and further study of this issue appears to be warranted.

The findings from experiment 3 showed that listeners were able to discriminate the relative strength of the statistical dependency present in two successive sound sequences significantly more accurately than in two concurrent sound sequences. This was true both for synchronous and asynchronous conditions. However, the difference between successive and asynchronous-concurrent presentation was smaller

than for synchronous-concurrent presentation, consistent with the proposition that the synchronous presentation of the two concurrent sequences promoted the perception of temporal coherence of the sounds, making the separate evaluation of the two streams more difficult. Our interpretation of this finding is broadly consistent with that of [Micheyl et al. \(2013\)](#) who concluded that a strict interpretation of temporal coherence theory (e.g., [Shamma et al., 2010](#)) could not account for their findings. In the [Micheyl et al. \(2013\)](#) study, they measured the discriminability of a slight alteration in frequency of the final element in a sequence of tones embedded in a context of other tones. The context of the embedded tone sequence was varied in certain ways (e.g., harmonically related or not to the target sequence) that either promoted across-frequency grouping or worked in opposition to grouping. When the context and target sequence were grouped together into a common stream due to harmonicity (and synchronous onsets), discriminating the details of the target sequence—the alteration in frequency of one element—was more difficult than when the target sequence was segregated. This segregation occurred because of the inharmonicity of the target sequence and context despite synchronous gating of the tones. Thus, as in the present study, temporal coherence appears to be a strong—but not always overriding—factor promoting the grouping of sound elements. These findings also are consistent with the idea that grouping/segregation is a perceptual continuum and not always a dichotomous one.

In the present study, it is not certain how the listeners solved the task in the concurrent case or the reasons underlying the performance decrement relative to successive presentation. One likely possibility is that the poorer performance in the concurrent case occurred because of the need for the listener to switch attention between the two sequences during the observation interval. In that case, the listener could process each stream while attended but would miss the information in the other stream or subsequently depend on whatever stored representation in memory of the unattended stream was available. The poor quality of the memory store of the unattended stream might not prove viable for the task. Thus, only portions of each stream would be available for scrutiny with some undetermined amount of the information lost while switching occurred. This effectively could be considered similar to the decrement in performance that would be predicted for sequences comprising fewer elements.

A second related possibility is that the listener divided processing resources between the two sequences, evaluating both simultaneously during the observation interval. This would seem to be a daunting task. Discriminating between two sequences varying in transition probability inherently involves both perceiving that transitions occur—which depends on the way that the states are implemented acoustically and the fidelity of the sensory representations of the stimuli—and in some manner recognizing how numerous the transitions are. Furthermore, explicit counting or subitizing two concurrent streams of events would be unlikely to yield the levels of performance found here given how sensitive such mental processes are to disruption by competition or divided attention (e.g., [Buchner et al., 1998](#); [Anobile et al., 2012](#)). Thus, although there was a decrement in

performance for concurrent presentation in both the synchronous and asynchronous conditions of experiment 3, the notion of achieving accurate counts from two simultaneous, rapid streams of sounds seems implausible.

A third possible explanation for the difference in performance between successive and concurrent presentation is that the listener fully attended only to one of the two sequences and used a strategy based on all of the transition matrices tested in the experiment. That is, a criterion was established at some point between the different distributions and the listener chose the attended stimulus when it had fewer transitions than the criterion or chose the other stimulus when it did not. This essentially would then become a YES-NO task. However, although the composite distribution formed from all of the transition matrices is rather broad, simulations suggest that this is a viable strategy yielding above-chance performance.⁶

The manipulations tested here were not intended to distinguish among these alternative explanations. In a sense, however, the issues raised by the decrement in performance due to concurrent presentation can be considered an instance of the general dual task with separate processing and decisions made for each stimulus (cf., [Navon and Gopher, 1979](#); [Bonnell and Hafter, 1998](#); [Gallun et al., 2007](#)). Performance in similar conditions requiring judgments of complex stimuli have been reported under conditions of high uncertainty with large costs often associated with the dual-task (e.g., [Best et al., 2006](#); [Ihlefeld and Shinn-Cunningham, 2008](#)).

Because the sequences were separated both by ear and by frequency region, it was possible for the listener to use either distinction as a means of identifying the target. When the sequences were presented successively, it made no difference which type of report was required or whether that was known before the trial. However, performance was worse (by six percentage points, on average) in the concurrent presentation condition for report by frequency band. We interpret this relatively small detrimental effect as being due to competition—at some level of processing—between evaluating the stimulus sequence for a dependency in frequency *state* and the need to identify the frequency *band* where the stronger dependency occurred. This may be because of the need to make two decisions/take actions along the same dimension, i.e., frequency, with a substantial difference in scale. That is, the difference in frequency states forming the sequential dependency was a fine distinction requiring the focus of attention on small frequency differences. The judgment about which sequence was the target was made along the same dimension but at a much coarser level—high band vs low band. The adverse effect of response type was apparent when the sequences were presented concurrently and performance was relatively poor. Perhaps this finding indicates that the within-dimension competition only occurs when the task demands are relatively high as is the case when concurrent streams must be evaluated.

B. Modeling human performance by degrading Ideal Observer performance

An important first point to make is that Ideal Observer performance does not depend on the particular stimulus

dimension from which the states are defined.⁷ Thus, the same predictions apply equally for the conditions in the experiments in which frequency and/or ITD values form the states. These computed levels of performance—determined solely by the underlying transition matrices—would apply if the samples to be compared on a trial were drawn from different stimulus dimensions or even if the set of states comprised unrelated sounds such as a noise burst, a chime, a vowel sound, etc. that were presented as sequences of events. Under the conditions tested here, human performance was poorer than ideal by amounts that varied according to which stimulus dimension comprised the states in the Markov chains, frequency or ITD, the acoustical properties of the stimulus (e.g., tones vs noise for location), and the values specified in the transition matrices. Although rate/duration was not a significant factor in the conditions tested here, it seems likely that performance could also be affected by more extreme ranges of values. Moreover, as suggested in the discussion of experiment 1 where performance was examined as a function of the number of trials, learning the properties of the stimulus distributions is key to solving the task.⁸ A comprehensive model of human performance would be able to describe the role and predict the effects of each of these factors. However, such a model is beyond the scope of the current work and it will require considerably more research before reaching that point. It is possible, though, to begin work toward the goal of developing such a model by taking the Ideal Observer as a starting point and then degrading performance according to the observations about human limitations in performing these experimental tasks. Note that these degradations are not necessarily grounded in insights into the constraints imposed by natural sound fields but rather are more limited and address only a few of the factors that may play a role in these laboratory-based conditions.

Figure 8 illustrates the performance of two models that degrade performance relative to that of the Ideal Observer in different ways. In this figure, performance is plotted in percent correct as a function of the difference in transition probability in a 2I-2AFC task (as used in experiments 2 and 3). Ideal Observer performance is shown as the dashed-dotted line. The performance functions for two different models are also shown (data points, see symbol key) along with the results from the human observers in one condition of experiment 2 (dashed line). The first model [labeled “Ideal ($n=4$)”] uses an optimal decision rule, that is, it bases a decision on the likelihood ratio, but assumes that the listener is only able to operate on a subset of the elements present in the sequence. This is equivalent to evaluating a sequence comprising fewer elements. There are various ways that this could correspond to human performance. Because the ability of an observer to operate on sequences of sounds inherently requires some form of memory or storage of past events (e.g., [van Noorden, 1975](#); [Jones, 1976](#); [Denham and Winkler, 2006](#); [Winkler et al., 2012](#); reviews in [Bregman, 1990](#); [Demany and Semal, 2008](#)), performance would reasonably be expected to depend on the size and viability of the storage mechanisms available. Thus, memory capacity could effectively limit the length of the sequence that could be processed in arriving at a decision. Or, in cases where multiple streams must be evaluated

FIG. 8. A comparison of a subset of the results from experiment 2 (dashed lines; target pitch with location constant) with three models (refer to text): Ideal Observer performance (dashed-dotted line) operating on the entire 12-element sequence, limited sequence-length model (diamonds) and longest-chain model (squares). The abscissa is the difference in transition probability between the matrices from which the target and non-target sequences were drawn. The ordinate is percent correct discrimination. Chance performance is 50% correct.

concurrently, it is plausible that only a subset of the elements in any sequence would be available to use in formulating a decision. That is, only the elements under the focus of attention are stored and subsequently available for processing with some loss of information occurring due to switching attentional focus. The limited-length model illustrated in Fig. 8 was restricted to sequences of length 4. This value was chosen simply because it fit one experimental condition rather well (the pitch dimension tested in experiment 2). The results for the target dimension of pitch with a fixed secondary dimension (location) reflect the best human performance found in the 2I-2AFC task.

Figure 9 is a plot of Ideal Observer performance for the limited sequence-length model. The left panel of Fig. 9 contains plots of percent correct performance as a function of the difference in the target and non-target transition matrices. The parameter of the graph is the number of elements comprising the sequences. The right panel is a plot of the difference in transition probability corresponding to 75% correct performance (an arbitrary designation of “threshold”) as a function of sequence length taken from the functions in the left panel. All of the functions originate at chance for no difference in transition probability then reach asymptote at different points. Consequently, because of the different slopes of the functions in the left panel, the “thresholds” depend on the percent correct point designated and, for 75% correct, decrease rapidly as the length of the sequence increases. The idea that the processing of sequences by searching for patterns among the elements is capacity limited makes some intuitive sense, so this model would appear to have some appeal on a theoretical level.

FIG. 9. The variation in Ideal Observer performance given the limited-length model (see Fig. 8 and text). The left panel shows percent correct performance in a 2I-2AFC task as a function of the difference in the transition matrices from which the stimuli were drawn. The parameter is the length of the sequence (number of elements drawn). The right panel contains the values along the abscissa corresponding to 75% correct performance for each of the functions shown in the left panel; i.e., discrimination “thresholds” as a function of the length of the sequence for the Ideal Observer.

The second model for which predictions are shown in Fig. 8 employs a non-optimal decision rule based on a “counting strategy” (labeled “longest chain”). Rather than basing a decision on the likelihood ratio, this model corresponds to performance that would occur in a 2I-2AFC task assuming that the listener chooses the stimulus on each trial that had the longest single chain of consecutive repetitions of state within a sequence (i.e., the longest string of elements without a transition). Such an approach would often be correct although it performs less well than ideal. However, on a practical level a counting strategy does not seem plausible given the lack of effect of rate/duration observed in experiment 1 or the relatively small decrement in concurrent performance in experiment 3. The way that human listeners would judge numerosity under the current conditions is not clear. There is a long history of study of judgments of the numerosity of stimuli—largely for visual stimuli but to a lesser degree for auditory stimuli (e.g., Garner, 1951; Pollack, 1968) and for the interactions among the modalities—and the memory processes that underlie such judgments. Despite the fact that an explicit counting strategy theoretically could yield performance comparable to that found here (cf. Figure 8), successfully implementing such a strategy by human observers seems unlikely. The process of explicit counting is complex and requires several mental operations (e.g., Logie and Baddeley, 1987) involving computation and storage in memory. Even the automatic—possibly preattentive—estimation of counts for small numbers of auditory events (“subitizing”; cf. Repp, 2007; Camos and Tillman, 2008; Pincham and Szucs, 2012) does not appear to provide a plausible account of the findings here. The rapid presentation rate of the stimuli employed here, combined with the invariance of performance with duration/rate in experiment 1, does not support such an interpretation. However, some listeners reported that they had experimented with such an approach under some conditions. Both of these non-ideal models, and likely others we did not test,

can be adjusted to yield performance approximating that of our listeners. The extent to which these or other models capture actual listener constraints and strategies, though, is an open question that will require further examination in the future.

VII. CONCLUSIONS

Human listeners are able to reliably detect experimentally defined statistical dependencies in sequences of sounds. Moreover, they can discriminate differences in the degree to which sequences of sounds exhibit statistical dependencies. This ability was apparent when the dependency, which was implemented by generating sequences of sound events using Markov chains, was implemented along the perceptual dimensions of pitch and apparent interaural location. For the values tested, but not necessarily for different ranges of values, detecting or discriminating a statistical dependency was better when the states were represented along the stimulus dimension of frequency than ITD. Discrimination performance was adversely affected by random variation along an irrelevant stimulus dimension and was markedly poorer when the test sequences were presented concurrently than successively. The difference between concurrent and successive presentation was reduced somewhat when the concurrent sequences were asynchronous, suggesting that temporal coherence can play a role in discrimination.

Human listener performance was poorer than Ideal Observer performance in all cases, but more so under some conditions than others (e.g., location of tones in experiment 2). Because Ideal Observer performance depends on the transition matrix, the variations in human performance due to the precise experimental conditions, e.g., the stimulus dimension representing states, thus reveal limitations on human processing. Some preliminary modeling efforts aimed at capturing these limitations degraded Ideal Observer performance by limiting the length of the Markov chains

available for processing and by implementing non-optimal decision rules.

These results reveal aspects of auditory stream formation that rely on the predictable variation of the elements comprising sequences of sounds. This approach appears to be a feasible means for studying stream formation and maintenance over time and under various types of adverse influences.

ACKNOWLEDGMENTS

Work supported by AFOSR grant FA9550-12-1-0171 and NIH/NIDCD grants DC04545, DC00100 and DC04663.

¹It could certainly be argued that a quasi-continuous, steady-state sound (e.g., a warning siren) forms a “stream” just as strongly as does a sequence of discrete, related elements. For real listeners, of course, even “continuous” sounds have a beginning and an end and fluctuate perceptually as the focus of attention waxes and wanes according to circumstance.

²van Noorden (1975) states that “coherence” is the “...relation between successive tones in which the observer has the impression that the sequence in question forms a whole which is ordered in time...” (p. 12 Glossary of Terms).

³Imagine the task of scientists monitoring signals received from space who are attempting to determine whether a particular source reveals the presence of intelligent life (e.g., SETI). What sort of signal would be taken as conclusive evidence? One plausible case might be when the signal consisted of a non-random sequence of states or separate elements that appear to exhibit some form of dependency over time. A listener in a highly uncertain multisource environment presumably implements segregating and selecting sound sources based on an analogous principle: Meaningful sources exhibit non-random sequential dependencies and deserve our scrutiny. Note that we are not claiming that a Markov model is apropos as a general model of natural sound sources but rather that the ability to distinguish orderly from random sequences of sound events is inherent to natural sound perception.

⁴Given a random sequence of states, $X^{(n)} = \{x_0, x_1, \dots, x_n\}$, over a finite set of states, $S = \{s_1, s_2, \dots, s_m\}$, and a transition matrix T that defines the conditional probability of transitioning from s_i to s_j , with elements $T_{ij} = \Pr(x_k = s_j | x_{k-1} = s_i)$, and a probability on the initial state sequence, $p_i^{(0)} = \Pr(x_0 = s_i)$, the likelihood that the Markov sequence was generated according to T and $p^{(0)}$ is $L(X^{(n)}/T, p^{(0)}) = p_i^{(0)} \prod_{k=1}^n T_{i_k, i_{k-1}}$. For a Markov chain where every current state is statistically independent of all past states, the transition matrix is uniform with $T_{ij} = 1/m$, and the likelihood of any sequence is $L(X^{(n)}/T, p^{(0)}) = p_i^{(0)} (1/m)^n$. The likelihood ratio test, used by an optimal receiver, for deciding whether a sequence of states $X^{(n)}$ was generated by a given system described by T or by one that is statistically independent is given by $L(X^{(n)}/T, p^{(0)}) > p_i^{(0)} (1/m)^n$ under the assumption that either system is equally likely to have generated the sequence.

⁵Transition probabilities greater than that associated with the non-target matrix (uniform, about 0.83) would yield cases where the “target” sequences would be expected to have more transitions than the non-target sequences. In the limit, for a transition probability of 1.0, a transition would always occur between adjacent events in the non-target sequence while fewer transitions would be expected given the uniform non-target matrix. Although we did not measure human performance under those conditions, which could produce ambiguous cues to detection, the resulting Ideal Observer performance function is, consequently, non-monotonic, with a minimum (chance) at about 0.83 and better-than-chance performance on either side of that value since the likelihood ratios would not be unity.

⁶Simulated results based on a composite distribution formed by all of the matrices used in experiment 3 assuming listeners only attended to one stream and made a YES-NO decision relative to a fixed and optimally placed criterion.

⁷Ideal Observer performance is based purely on the likelihood ratio formed by the underlying distributions of target and nontarget stimuli and thus does not capture many of the constraints likely encountered by human observers in natural sound fields.

⁸In order to successfully judge samples from target and non-target distributions, the listeners presumably were required to learn, to a sufficient

degree, the characteristics of these distributions and, given the YES-NO task, could solve the task by knowledge of either or both types of distributions. The subjects appeared to learn this task quickly and were able to achieve performance well above chance within the first few blocks of trials. Statistical learning normally occurs in many everyday circumstances beginning as early in life as the acquisition of language by infants (e.g., Saffran *et al.*, 1996), and it seems likely that this process is integral to the way in which listeners form the associations that bind sound elements into streams.

Anobile, G., Turi, M., Cicchini, G. M., and Burr, D. C. (2012). “The effects of cross-sensory attentional demand on subitizing and on mapping number onto space,” *Vision Res.* **74**, 102–109.

Berliner, J. E., and Durlach, N. I. (1973). “Intensity perception. IV. Resolution in roving-level discrimination,” *J. Acoust. Soc. Am.* **53**, 1270–1287.

Best, V., Gallun, F. J., Ihlefeld, A., and Shinn-Cunningham, B. G. (2006). “The influence of spatial separation on divided listening,” *J. Acoust. Soc. Am.* **120**, 1506–1516.

Bonnel, A. M., and Hafer, E. R. (1998). “Divided attention between simultaneous auditory and visual signals,” *Percept. Psychophys.* **60**, 179–190.

Braida, L. D., and Durlach, N. I. (1972). “Intensity perception II. Resolution in one-interval paradigms,” *J. Acoust. Soc. Am.* **51**, 483–502.

Bregman, A. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sounds* (MIT Press, Cambridge, MA).

Bregman, A. S. (2008). “Auditory scene analysis,” in *The Senses: A Comprehensive Reference, Vol. 3: Audition*, edited by A. I. Basbaum, A. Koneko, G. M. Shepherd, and G. Westheimer (Academic Press, San Diego, CA), pp. 861–870.

Brouwer, S., Van Engen, K. J., Calandruccio, L., and Bradlow, A. R. (2012). “Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content,” *J. Acoust. Soc. Am.* **131**, 1449–1464.

Buchner, A., Steffens, M. C., Irmen, L., and Wender, K. F. (1998). “Irrelevant auditory material affects counting,” *Cognition* **24**, 48–67.

Camos, V., and Tillman, B. (2008). “Discontinuity in the enumeration of sequentially presented auditory and visual stimuli,” *Cognition* **107**, 1135–1143.

Culling, J. F., and Colburn, H. S. (2000). “Binaural sluggishness in the perception of tone sequences and speech in noise,” *J. Acoust. Soc. Am.* **107**, 517–527.

Darwin, C. J., and Carlyon, R. P. (1995). “Auditory grouping,” in *Hearing*, edited by B. C. J. Moore (Academic Press, San Diego, CA), pp. 387–424.

Demany, L., and Semal, C. (2008). “The role of memory in auditory perception,” in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer, New York), pp. 77–113.

Denham, S. L., and Winkler, I. (2006). “The role of predictive models in the formation of auditory streams,” *J. Physiol. Paris* **100**, 154–170.

Durlach, N. I. and Colburn, H. S. (1978). “Binaural phenomena,” in *Handbook of Perception*, edited by E. C. Carterette and M. P. Friedman (Academic Press, San Diego, CA), Vol. 4, pp. 365–466.

French-St. George, M., and Bregman, A. S. (1989). “Role of predictability of sequence in auditory stream segregation,” *Percept. Psychophys.* **46**, 384–386.

Gallun, F. J., Mason, C. R., and Kidd, G., Jr. (2007). “Task-dependent costs in processing two simultaneous auditory stimuli” *Percept. Psychophys.* **69**, 757–771.

Garner, W. R. (1951). “The accuracy of counting repeated short tones,” *J. Exp. Psychol.* **41**, 310–316.

Hall, J. W., III, Buss, E., and Grose, J. H. (2011). “Masked detection and discrimination of tone sequences under conditions of monaural and binaural masking release,” *J. Acoust. Soc. Am.* **129**, 1482–1489.

Harris, J. D. (1952). “The decline of pitch discrimination with time,” *J. Exp. Psych.* **43**, 96–99.

Helfer, K. S., Chevalier, J., and Freyman, R. L. (2010). “Aging, spatial cues, and single-versus dual-task performance in competing speech perception,” *J. Acoust. Soc. Am.* **128**, 3625–3633.

Huang, R., and Richards, V. M. (2008). “Estimating internal templates for the detection of sequential tonal patterns,” *J. Acoust. Soc. Am.* **124**, 3831–3840.

Ihlefeld, A., and Shinn-Cunningham, B. G. (2008). “Spatial release from energetic and informational masking in a divided speech identification task,” *J. Acoust. Soc. Am.* **123**, 4380–4392.

- Jones, M. (1976). "Time, our lost dimension: Toward a new theory of perception, attention and memory," *Psychol. Rev.* **83**, 323–355.
- Kidd, G., Jr., Mason, C. R., and Arbogast, T. L. (2002). "Similarity, uncertainty and masking in the identification of nonspeech auditory patterns," *J. Acoust. Soc. Am.* **111**, 1367–1376.
- Kidd, G., Jr., Mason, C. R., and Dai, H. (1995). "Discriminating coherence in spectro-temporal patterns," *J. Acoust. Soc. Am.* **97**, 3782–3790.
- Kidd, G., Jr., Mason, C. R., and Hanna, T. E. (1988). "Evidence for sensory-trace comparisons in spectral shape discrimination," *J. Acoust. Soc. Am.* **84**, 144–149.
- Kidd, G., Jr., Mason, C. R., and Richards, V. M. (2003). "Multiple bursts, multiple looks and stream coherence in the release from informational masking," *J. Acoust. Soc. Am.* **114**, 2835–2845.
- Kidd, G., Jr., Streeter, T., Thompson, E., Mason, C. R., Best, V., and Wakefield, G. (2012). "The role of listener expectation in the formation and maintenance of auditory streams," Program of the 35th Annual Meeting of the Association for Research in Otolaryngology, San Diego, CA, p. 39.
- Klumpp, R. G., and Eady, H. R. (1956). "Some measurements of interaural time-difference thresholds," *J. Acoust. Soc. Am.* **28**, 859–860.
- Licklider, J. C. R. (1952). "On the process of speech perception," *J. Acoust. Soc. Am.* **24**, 590–594.
- Logie, R. H., and Baddeley, A. D. (1987). "Cognitive processes in counting," *J. Exp. Psychol.* **13**, 310–326.
- Lutfi, R. A. (2008). "Human sound source identification," in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer, New York), pp. 13–42.
- Maddox, R. K., and Shinn-Cunningham, B. G. (2012). "Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention," *J. Assoc. Res. Otolaryn.* **13**, 119–129.
- Micheyl, C., Carlyon, R. P., Gutschalk, A., Melcher, J. R., Oxenham, A. J., Rauschecker, J. P., Tian, B., and Wilson, C. (2007). "The role of auditory cortex in the formation of auditory streams," *Hear. Res.* **229**, 116–131.
- Micheyl, C., Kreft, H., Shamma, S., and Oxenham, A. J. (2013). "Temporal coherence versus harmonicity in auditory stream formation," *J. Acoust. Soc. Am.* **133**, EL188–EL194.
- Mill, R., Coath, M., Wennekers, T., and Denham, S. L. (2011). "A neuro-computational model of stimulus-specific adaptation to oddball and Markov sequences," *PLoS Comput Biol.* **7**(8), e1002117.
- Moore, B. C. J., and Gockel, H. (2012). "Properties of auditory stream formation," *Philos. Trans. R. Soc. B.* **367**, 919–931.
- Navon, G., and Gopher, D. (1979). "On the economy of the human processing system," *Psychol. Rev.* **86**, 214–255.
- Overath, T., Kumar, S., Stewart, L., van Kriegstein, K., Cusack, R., Rees, A., and Griffiths, T. D. (2010). "Cortical mechanisms for the segregation and representation of acoustic textures," *J. Neurosci.* **30**, 2070–2076.
- Pincham, H. L., and Szucs, D. (2012). "Intentional subitizing: Exploring the role of automaticity in enumeration," *Cognition* **124**, 107–116.
- Pollack, I. (1968). "Number or duration discrimination for auditory pulse trains," *J. Exp. Psychol.* **78**, 660–665.
- Pollack, I. (1970). "Depth of sequential auditory information processing. II," *J. Acoust. Soc. Am.* **48**, 906–912.
- Rabiner, L. R. (1989). "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceed. IEEE* **77**, 257–286.
- Repp, B. H. (2007). "Perceiving the numerosity of rapidly occurring auditory events in metrical and nonmetrical contexts," *Percept. Psychophys.* **69**, 529–543.
- Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). "Statistical learning by 8-monthold infants," *Science* **274**, 1926–1928.
- Shamma, S. A., Elhilali, M., Micheyl, C. (2010). "Temporal coherence and attention in auditory scene analysis," *Trends Neurosci.* **34**, 114–123.
- van Noorden, L. P. A. S. (1975). "Temporal coherence in the perception of tone sequences," Ph.D. dissertation, Eindhoven University of Technology, Eindhoven, Netherlands.
- Watson, C. S. (2005). "Some comments on informational masking," *Acta Acust. Acust.* **91**, 502–512.
- Winkler, I., Denham, S., Mill, R., Bohm, T. M., and Bendixen, A. (2012). "Multistability in auditory stream segregation: A predictive coding view," *Proc. R. Soc. B. Biol. Sci.* **367**(1591), 1001–1012.
- Yost, W. A. (1991). "Auditory image perception and analysis: The basis for hearing," *Hear. Res.* **56**, 8–18.
- Yost, W. A. (2008). "Perceiving sound sources," in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer, New York), pp. 1–13.
- Yost, W. A., Braida, L. D., Hartmann, W. M., Kidd, G. Jr., Kruskal, J. B., Pastore, R. E., Sorkin, R. D. and Warren, R. M. (1989). *Classification of Complex Nonspeech Sounds*, Committee on Hearing, Bioacoustics and Biomechanics, National Research Council, National Academy Press, Washington, D.C., 88 pages.