Perceptual–Gestural (Mis)Mapping in Serial Short-Term Memory: The Impact of Talker Variability

Robert W. Hughes, John E. Marsh, and Dylan M. Jones
Cardiff University

The mechanisms underlying the poorer serial recall of talker-variable lists (e.g., alternating female–male voices) as compared with single-voice lists were examined. We tested the novel hypothesis that this talker variability effect arises from the tendency for perceptual organization to partition the list into streams based on voice such that the representation of order maps poorly onto the formation of a gestural sequence-output plan assembled in support of the reproduction of the true temporal order of the items. In line with the hypothesis, (a) the presence of a spoken lead-in designed to further promote by-voice perceptual partitioning accentuates the effect (Experiments 1 and 2); (b) the impairment is larger the greater the acoustic coherence is between nonadjacent items: Alternating-voice lists are more poorly recalled than four-voice lists (Experiment 3); and (c) talker variability combines nonadditively with phonological similarity, consistent with the view that both variables disrupt sequence output planning (Experiment 4). The results support the view that serial short-term memory performance reflects the action of sequencing processes embodied within general-purpose perceptual input-processing and gestural output-planning systems.

Keywords: short-term memory, talker variability, serial recall, perceptual–gestural view, embodied cognition

One major class of accounts of serial short-term memory is centered at the item level and assumes that explanations of serial behavior will flow from an understanding of item-level properties, such as the rate of item decay or/and the structural (e.g., phonological) similarity of one item to another (e.g., Baddeley, 1986, 2007; Farrell & Lewandowsky, 2002; Nairne, 1990; Neath, 2000). However, an alternative framework—the perceptual–gestural account—focuses on factors that operate at a level superordinate to the item, at the level of sequence formation, both at input (particularly in the auditory modality, in the formation of streams) and at motor output planning in the formation of a sequence of subvocal gestures (Hughes & Jones, 2005; Jones, Hughes, & Macken, 2006, 2007; Jones, Macken, & Nicholls, 2004; Woodward, Macken, & Jones, 2008). In the present article, we sought to illustrate the importance of sequence-level factors in understanding serial short-term memory performance by capitalizing on the disruptive impact on auditory-verbal serial recall of presenting lists in more than one voice, particularly alternating female–male voices (Greene, 1991). Evidence is presented that this talker variability effect (TVE) results from the fact that obligatory formation of voice-based streams (auditory streaming; see Bregman, 1990) produces a poor mapping between the perception of order (within streams) and the need to assemble the items into a gestural (articulatory) sequence output plan that mimics the canonical order of the items.

Serial Short-Term Memory: A Perceptual–Gestural Account

Current understanding of serial verbal short-term memory is based largely on the serial recall paradigm in which a list of verbal items (e.g., digits, letters, words) must be recalled in strict serial order (Baddeley, 1966; Conrad, 1964). Classically, explanations of serial recall have tended to focus on the properties of the individual list items. For instance, according to the decay-rehearsal approach to verbal short-term memory (e.g., the phonological loop model; Baddeley, 1986, 2007; Baddeley & Hitch, 1974; see also Cowan, 1995), verbal items enter a passive, bespoke, phonological store dedicated to the temporary retention of verbal events (Baddeley, 1986, 2007). Items in the store are subject to decay within a few seconds unless refreshed by a separate articulatory control process and are also susceptible to mutual interference by virtue of their structural (e.g., phonological) similarity (e.g., Baddeley, 1986). Other accounts posit that similarity-based item interference can account for serial recall phenomena without the notion of item decay (e.g., Lewandowsky, Geiger, & Oberauer, 2008; Nairne,
An alternative view—the perceptual–gestural account—holds that serial recall performance is parasitic on general-purpose perceptual and motor-planning processes that operate at the level of the sequence, not each item (e.g., Hughes & Jones, 2005; Jones et al., 2006, 2007, 2004; Woodward et al., 2008).

An important feature of the typical serial recall study is that the burden of processing lies with reproducing the order of the items: A familiar closed set of items is typically used on each trial (e.g., permutations of 1–8), and hence the individual items are known before presentation (hereafter: pure serial recall). The starting point for the perceptual–gestural view is the characterization of the typical serial recall list as a sequence in which the intrinsic transitional probabilities between successive items—the predictability of an event given the preceding event(s) (e.g., Miller & Chomsky, 1963)—are, by design, very low: Syntax, grammar, and semantics that in natural language constrain temporal order are stripped from the serial recall list (see Jefferies, Lambon Ralph, & Baddeley, 2004; Macken & Jones, 2003). It is this feature that makes serial recall difficult. For example, performance is superior when there is a good match between the list and long-term sequential knowledge, such as for lists containing high-frequency letter transitions (Miller & Selfridge, 1950), lists of words that make up a grammatically legitimate sentence (Jefferies et al., 2004), or lists of adjective–noun pairings compared with less frequently encountered (in English) noun–adjective pairings (Perham, Marsh, & Jones, 2009).

In the absence of strong order cues in the typical serial recall list, how is serial order preserved? We argue that transitional probabilities are crafted onto the material by co-opting the generic skill of gestural sequence planning (vocal articulatory in the case of verbal material; Macken & Jones, 2003). The skill of speech planning lends itself well to this task because of its inherent sequentiality and its range of paralinguistic sequencing subskills, such as prosody and coarticulation (Stenberg, Wright, Knoll, & Monsell, 1980). For example, prosodic characteristics of timing and intonation deployed in natural phrase and sentence production may be brought to bear to minimize transitional probabilities across group boundaries. This provides strong cues to order at boundaries that constrain migration of items across groups (e.g., Maybery, Parmentier, & Jones, 2002). Moreover, the skill of coarticulation—adjusting the way in which the end of one speech element (e.g., syllable, word) is articulated so as to lend fluency to the transition to the next element (Stenberg et al., 1980)—also serves to increase transitional probabilities between successive items (as shown by the finding that serial recall is a positive function of coarticulatory fluency; Murray & Jones, 2002; Woodward et al., 2008). Thus, an articulatory plan is assembled as a surrogate for the lack of correspondence between long-term sequence knowledge and the to-be-reproduced sequence; the motor plan is the very agent by which item order is realized and supported. Explanations of serial recall phenomena are sought, therefore, by recourse to factors that emerge at the level of the articulated sequence and that are not, by definition, to be found at the local item level.

The assembly of the articulatory plan is time critical. In their presentation, sequences are paced and the items evanescent. The particular items need to be loaded successfully and in a timely and orderly fashion onto the necessarily abstract and generic motor sequence plan—a process involving a number of conversion operations that could include transformation from graphic form, coarticulation, and so forth. Thus, numerous factors, such as spatial or temporal uncertainty during presentation or the complexity of the gestures required to (co)articulate successive items, can potentially render the loading process more fraught and hence compromise serial recall performance (e.g., Murray & Jones, 2002). Of particular relevance in the present article, when items are presented auditorily, powerful processes of perceptual organization may also influence the process of picking up and embodying elements in the plan in their correct order.

The perceptual component of the perceptual–gestural account is based largely on Bregman’s (1990) auditory scene analysis framework: This refers to the preattentive and obligatory (i.e., nonvolitional) partitioning of the mixture of acoustic signals reaching the ears into discrete mental descriptions (streams) corresponding to each individual sound-emitting event contributing to that mixture (for evidence for the obligatory nature of auditory streaming, see, e.g., Deouell, Deutsch, Scabini, Soroker, & Knight, 2008; Macken, Tremblay, Houghton, Nicholls, & Jones, 2003). An aspect of streaming with particular relevance for serial recall is sequential streaming: the computation of whether temporally successive auditory stimuli share a common origin, a task accomplished by exploiting Gestalt-like grouping principles such as similarity of frequency, timbre, and good continuation (see Bregman, 1990, Chapter 2; Warren, 1999). Thus, successive events with a relatively low acoustic-level transitional probability—such as would be the case with the alternation of two tones highly distinct in pitch (or fundamental frequency)—are relatively unlikely to be computed as having the same origin and hence will tend to be partitioned to form two distinct streams (e.g., Miller & Heise, 1950; Rogers & Bregman, 1993; van Noorden, 1975). In such cases, “the auditory system is grouping tones that are similar to one another in preference to grouping tones that follow one another immediately in time” (Bregman, 1990, p. 45). In the present article, we sought to demonstrate how such perceptual grouping of elements by spectral similarity may influence—in this case impair—the formation of the articulatory plan assembled in the service of order retention by examining the disruptive impact of talker variability on serial recall (e.g., Greene, 1991).

The TVE in Serial Recall

Auditory-verbal serial recall is impaired if the list is conveyed in more than one voice (e.g., Goldinger, Pisoni, & Logan, 1991; Greene, 1991; Martin, Mullenix, Pisoni, & Summers, 1989; Nygaard, Sommers, & Pisoni, 1995). Previous accounts of this TVE in serial recall1 have appealed to essentially the same explanation that the decay-rehearsal model offers for the word-length effect (e.g., Baddeley, Thomson, & Buchanan, 1975): The increased time taken to encode or/and rehearse talker-variable items—just as with long compared with short words—impairs recall by delaying the opportunity to refresh decay-prone items held in a bespoke pho-

---

1 TVEs have been found in several domains other than short-term memory (long-term word recognition; e.g., Mullenix, Pisoni, & Martin, 1989). However, in the present article we are concerned specifically with the putative mechanisms underpinning the impact of talker variability in serial recall, and we use the acronym TVE in this restricted sense unless indicated otherwise (also see the General Discussion section).
nological store. On one such account, the item-refresh delay results from talker-variable lists imposing a greater burden on talker normalization (cf. Pisoni, 1997), whereby extra-linguistic, indexical properties such as the accent, gender, speaking style, and emotional state of the particular speaker (see Abercrombie, 1967) need to be discarded so as to yield abstract, canonical, and linguistic (i.e., phonological) item representations (Martin et al., 1989). In another account within this approach, the item-refresh delay is due to a process of incorporating the indexical properties rather than discarding them, a process that would again be under greater duress with more than one talker (Goldinger et al., 1991).

An alternative view based on the perceptual–gestural account is that the TVE reflects a mismapping between perceptual organization and the assembly of an articulatory plan. In a single-voice list, adjacent items are likely to be perceptually grouped together on account of their similarity in fundamental frequency and timbre as well as indexical properties such as accent (Pardo & Remez, 2006). In a talker-variable list, however, the transitional probabilities among immediately adjacent items—and hence the likelihood of the items being grouped together—may be greatly diminished. Indeed, as with alternating tones (e.g., Rogers & Bregman, 1993), when the same two voices (e.g., male and female) alternate in a list (Greene, 1991), the transitional probabilities are likely to be greater between items spoken within each voice, that is, between nonadjacent items. Importantly, several studies have shown that the perception of order is relatively good for a succession of events assigned to the same stream but is notoriously poor for events traversing different streams (e.g., Bregman & Campbell, 1971; Warren, Obusek, Farmer, & Warren, 1969). For example, if a sequence of high-frequency tones (A, B, C) is alternated with a sequence of low-frequency tones (1, 2, 3)—for example, A, 1, B, 2, C, 3—in a repeating loop, the majority of participants inadvertently report the order of the tones by frequency range (or stream; e.g., A, B, C, 1, 2, 3) rather than according to true temporal order (Bregman & Campbell, 1971; for similar effects using speech tokens, see Lackner & Goldstein, 1974). Such temporal order confusion—due this time to streaming by spatial location (e.g., Handel, 1989)—has also been used to explain the tendency to repeat back a sequence of simultaneous pairs of digits—one digit presented to each ear—by ear rather than by true temporal order (Broadbent & Gregory, 1961) and also of the difficulty of reproducing the true order of a sequence of ear-alternating items (Moray, 1960; Bregman, 1990; ten Hoopen, 1996). All these cases illustrate how streaming can yield a perception of order pertaining to acoustically proximal events that takes precedence over that for temporally proximal ones. In relation to the TVE, then, we hypothesize that, perceptually, the order of adjacent items in a single-voice list would be relatively well preserved. In contrast, with a talker-variable list, obligatory by-voice streaming produces a mismapping between perceived order and the deliberate attempt to impose articulatory-based transitional probabilities between adjacent items in support of strict serial recall.

We test the perceptual–gestural mismapping account in the present study by modulating in various ways the likely strength of by-voice perceptual partitioning of an alternating-voice list (Experiments 1–3). Our rationale was that if the TVE is driven by the perceptual incoherence of temporally adjacent items in an alternating-voice list, any factor that promotes that incoherence—that is, promotes by-voice streaming—should accentuate the TVE. We then go on in Experiment 4 to examine the contention that by-voice streaming has its disruptive impact by impairing the articulatory sequence-planning process.

Experiment 1

All experiments reported here follow Greene’s (1991) methodology and examine the impact of alternating female–male voices on strict serial recall of a closed set of verbal items (digits or letters). In Experiment 1, we put the perceptual–gestural mismapping account to the test by exploiting the fact that streaming is not an all-or-none phenomenon and that it, moreover, takes time to build up. We sought to accentuate the perceptual partitioning of alternating-voice lists by capitalizing on a form of auditory stream biasing whereby the likelihood of partitioning an alternating-stimulus sequence is strengthened if one of the stimuli in that sequence already forms part of a perceptually stable stream. More concretely, if a sequence of alternating high (H) and low (L) tones (HLHLHL) is preceded by a lead-in of either H or L tones (e.g., LLLLLHLHLHL), the partitioning of the eventual alternating sequence into two separate L and H streams occurs more readily. This is because the stable “LLLLL . . .” stream perceptually captures the L tones in the following alternating sequence, whereas the remaining H tones are more readily “thrown out” as “alien” tones to form their own distinct stream (Beauvois & Meddis, 1997; Rogers & Bregman, 1993). We have demonstrated elsewhere that the same principles hold also for speech stimuli (Nicholls & Jones, 2002; although see Remez, Rubin, Berns, Pardo, & Lang, 1994, for the view that for natural speech utterances there must also be domain-specific phonetic organization processes distinct from Gestalt-based ones). Thus, in this experiment, we sought to promote the perceptual partitioning of an alternating-voice list in a serial recall task by presenting a lead-in in the form of a countdown (“8, 7, 6 . . . 1”) spoken in the same rhythm as the ensuing to-be-remembered (TBR) items and spoken in just one of the two voices making up the ensuing alternating voice list.2

We also sought to promote by-voice partitioning of the TBR items by exploiting the fact that such partitioning takes some time to build up; it becomes more emphatic as the evidence that there are indeed two distinct acoustic events accumulates. Thus, “all stimuli [in an alternating-tone sequence] begin by sounding temporally coherent and . . . the probability of stream formation increases steadily over time as a function of sequence duration” (Beauvois & Meddis, 1997, p. 81; see also Anstis & Saida, 1985; Bregman, 1978, 1990; Carlyon, Cusack, Foxton, & Robertson, 2001). This build-up has a similar basis to the biasing effect described earlier: As an alternating sequence (LHLHLH . . .) continues, the ever-increasing stability of each individual stream—LLL and HHH—gradually increases the likelihood that each stream will capture each new L and H tone, respectively. Thus, we included a further condition that involved effectively increasing the duration of the alternating sequence by preceding an alternating-voice list with an alternating-voice lead-in. Again, this

---

2 The countdown did not, of course, have to be reproduced. The start of the list was marked by the end of the countdown, that is, once the countdown reached “1” (or “z” in Experiment 2), the participant would know that the next item was the first item of the TBR list.
type of lead-in should promote the partitioning of alternating-voice items in the TBR list—hence the magnitude of the TVE—because the by-voice partitioning process will have already begun during the lead-in and should be stronger therefore (compared with a no lead-in condition) by the time the TBR list starts. Table 1 provides a list of all six conditions contrasted in Experiment 1. Conditions 1 and 2 (i.e., those without a lead-in) are those required to show the standard TVE. The critical contrast thereafter will be between performance with an alternating-voice list (Alt) and that with an alternating-voice list preceded by a lead-in (Single-Alt or Alt-Alt): The perceptual–gestural mismapping account predicts poorer performance in the Single-Alt and Alt-Alt conditions compared with the Alt condition because the lead-in (of either type) should promote by-voice partitioning, thereby exacerbating the conflict between order perception and the articulatory assembly of the items in canonical order. Whereas the present experiment serves primarily as a test of our perceptual–gestural mismapping account, it is unclear how an item-decay-based approach (Goldinger et al., 1991; Martin et al., 1989) could accommodate any influence of lead-ins on the TVE.

Method

Participants. Twenty-two undergraduates from Cardiff University took part in return for course credits. Each participant reported normal hearing and normal or corrected-to-normal vision.

Apparatus and materials. The TBR lists comprised eight items taken without replacement from the digit set 1–8. Each item was recorded digitally once in a female voice and once in a male voice (the items within each voice were spoken at an approximately even pitch) and sampled with a 16-bit resolution at a sampling rate of 44.1KHz using Sound Forge 5 software (Sonic Inc., Madison, WI). The male and female voices clearly differed from one another on account of their distinct fundamental frequency and timbre. Each item’s duration was edited to 250 ms using the same software. For each TBR list, the digits were presented in a pseudorandom order with care taken to ensure that there were no more than two occasions across a given TBR list on which there was an ascending or descending run of two or more digits (e.g., 2–3 or 7–6) and that there were no runs of three or more digits. This was also the case for nonadjacent items (e.g., those in Positions 1 and 3) so that in alternating-voice lists there were no more than two two-digit runs within a given voice in a given list. The TBR list (and lead-in when present) was presented at approximately 65–70 dB(A) over stereo headphones with an interstimulus interval (ISI; offset to onset) of 100 ms, giving an item presentation rate of one item per 350 ms. The stimuli were presented using SuperLab software (Cedrus Corporation, San Pedro, CA).

Table 1 provides a schematic representation of the six conditions contrasted in Experiment 1. In the Alt-Single and Alt-Alt conditions, the lead-in was presented in alternating female–male fashion. In conditions involving a single-voice lead-in, a countdown was presented either in the same voice as the ensuing single-voice list (Single-Single) or, for the Single-Alt condition, in the same voice as that conveying the second, fourth, sixth, and eighth items of the ensuing alternating-voice list. In the Alt-Single and Alt-Alt conditions, the lead-in was presented in alternating female–male voices.

Design. The design involved three repeated measures factors: Lead-In (with three levels: no lead-in, single-voice lead-in, and alternating-voice lead-in), List-Type (with two levels: single-voice and alternating-voice), and Serial Position (eight levels). There were 84 experimental trials divided into two blocks: The with lead-ins block comprised 56 experimental trials made up of 14 Alt-Alt trials, 14 Alt-Single trials, 14 Single-Single trials, and 14 Single-Alt trials. The block was preceded by four practice trials, one from each of the four conditions. The other block—the without lead-ins block—comprised 28 experimental trials made up of 14 single-voice TBR lists and 14 alternating-voice TBR lists preceded by two practice trials, one from each condition. In both blocks, the various trial types were presented pseudorandomly with the constraint that no condition was presented more than twice in succession. The order in which the two blocks were undertaken was counterbalanced across participants. Further counterbalancing measures incorporated into Experiments 1–3 are provided in the Appendix.

Procedure. Participants were tested in groups of up to four in a sound-attenuated room with each participant placed in a separate cubicle with its own PC and headphones. Participants were to recall the TBR digits in their correct order and to ignore the

<table>
<thead>
<tr>
<th>Condition</th>
<th>Voice</th>
<th>Lead-in</th>
<th>To-be-remembered list</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Single</td>
<td>Female (or male) voice</td>
<td>6 5 2 7 1 4 8 3</td>
<td></td>
</tr>
<tr>
<td>2. Alt</td>
<td>Female (or male) voice</td>
<td>6 2 1 8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Male (or female) voice</td>
<td>5 7 4 3</td>
<td></td>
</tr>
<tr>
<td>3. Single-Single</td>
<td>Female (or male) voice</td>
<td>8 7 6 5 4 3 2 1</td>
<td>6 5 2 7 1 4 8 3</td>
</tr>
<tr>
<td>4. Alt-Alt</td>
<td>Female (or male) voice</td>
<td>8 6 4 2</td>
<td>5 7 4 3</td>
</tr>
<tr>
<td></td>
<td>Male (or female) voice</td>
<td>7 5 3 1</td>
<td></td>
</tr>
<tr>
<td>5. Alt-Single</td>
<td>Female (or male) voice</td>
<td>8 6 4 2</td>
<td>6 5 2 7 1 4 8 3</td>
</tr>
<tr>
<td></td>
<td>Male (or female) voice</td>
<td>7 5 3 1</td>
<td></td>
</tr>
<tr>
<td>6. Single-Alt</td>
<td>Female (or male) voice</td>
<td>8 7 6 5 4 3 2 1</td>
<td>5 7 4 3</td>
</tr>
<tr>
<td></td>
<td>Male (or female) voice</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. For Conditions 3–6, the first part of each condition label refers to the voice-presentation format of the lead-in, whereas the second part refers to that for the to-be-remembered list. Single = single voice; Alt = alternating voices.
particular voice(s) conveying the digits. Participants were also told that for one block of trials the spoken list would be preceded by a spoken countdown. They were informed that 100 ms following the offset of the last TBR item of each list, the cue “RECALL” would appear on the screen, at which point they were to write down the items in the correct order on response sheets marked with eight blank spaces for each trial. Participants had 15 s to recall the list and were instructed to do so in a strict left-to-right fashion such that they should start with Position 1, then Position 2, and so on. They were instructed to guess if they were uncertain of any of the digits’ positions. A 500-ms tone was presented over the headphones 13 s into the 15-s recall period to signal that the presentation of the first item of the next trial was imminent (in trials with a lead-in, the first item would be the first item of the countdown). The experiment lasted approximately 45 min.

Results and Discussion

For all experiments, the raw serial recall data were scored according to the strict serial recall criterion: To be recorded as correct, an item had to be recalled in its original presentation position. Figure 1 shows the percentage of items correctly recalled across the eight serial positions in the six conditions. The pattern of results is clear cut and can be unpacked initially into two distinct sets of curves: Replicating the basic TVE, performance in conditions involving an alternating TBR list (i.e., Alt, Single-Alt, and Alt-Alt; represented by the triangle symbols) was uniformly poorer than for conditions involving a single-voice list (i.e., Single, Alt-Single, and Single-Single; represented by the square symbols). More importantly, the TVE was markedly accentuated by the presence of a lead-in: Performance with alternating-voice lists was particularly poor when those lists were preceded by either lead-in (Single-Alt and Alt-Alt). The pattern across conditions thus conforms to that predicted by the perceptual–gestural mismapping account.

A 2 (List-Type) × 3 (Lead-In) × 8 (Serial Position) repeated measures analysis of variance (ANOVA) revealed a main effect of Serial Position, \( F(7, 147) = 55.27, MSE = 0.06, p < .001 \); a main effect of List-Type, \( F(1, 21) = 69.83, MSE = 0.07, p < .001 \); a main effect of Lead-In, \( F(2, 42) = 15.87, MSE = 0.01, p < .001 \); and, most importantly, a significant List-Type × Lead-In interaction, \( F(2, 42) = 12.19, MSE = 0.02, p < .001 \), reflecting the fact that the TVE was larger when an alternating list was preceded by a lead-in (of either type). The only other significant effect was an interaction between List-Type and Serial Position, \( F(7, 147) = 18.37, MSE = 0.01, p < .001 \), possibly reflecting ceiling effects at primacy and recency serving to obscure differences according to list-type. Follow-up simple effects analyses confirmed that all alternating-voice TBR list conditions produced poorer performance than any of the conditions with a single-voice TBR list (all comparisons, \( p < .005 \)). More importantly, they also showed that performance was poorer in both the Single-Alt and Alt-Alt conditions than in the Alt condition (both \( ps < .001 \)). A further diagnostically important feature of the data is that the presence of a lead-in per se had no effect on serial recall: There was no significant difference between either the Single-Single or Alt-Single and the Single condition (both comparisons, \( p > .05 \)).

The results of Experiment 1 confirm a prediction of a perceptual–gestural mismapping account of the TVE: The lead-in (of either type) promoted the perceptual incoherence of adjacent TBR items—and at the same time promoted the perceptual coherence of nonadjacent items—thereby accentuating the mismapping between the order suggested by streaming and the action requirements of the serial recall task. Accordingly, the TVE was significantly larger when an alternating-voice list was preceded by a lead-in. A possible alternative interpretation of these data, however, is that the particularly poor performance in the Alt-Alt and Single-Alt conditions was due to the digit lead-in producing pro-active interference (PI)—the difficulty of recalling items because of their postcategorical similarity to previously encountered items (e.g., Bunting, 2006; Underwood, 1957)—or producing some general attentional distraction effect (S. Lewandowsky, personal communication, December 1, 2008). This seems unlikely, however, because the presence of a lead-in had no effect when it preceded a single-voice list. There was no difference between the Alt-Single or Single-Single conditions and the Single condition. Nonetheless, it is possible to counterargue that a PI/distraction effect from a lead-in might only emerge if recall is already made difficult by having to recall an alternating-voice list. In Experiment 2, we

Figure 1. Mean percentage of items correctly recalled at each serial position in the six conditions of Experiment 1 (see Table 1 for an illustration of the six conditions). Single = single voice; Alt = alternating voices.

---

3 Given that the presentation rate used in Experiment 1 was relatively fast (one item per 350 ms), we ran a supplementary experiment to check that the same interaction between talker variability and lead-in is also found with a slower rate typical of some serial recall experiments (one item per 750 ms, e.g., Henson, Hartley, Burgess, Hitch, & Flude, 2003; Hughes, Vachon, & Jones, 2007). Other than the presentation rate—which we increased to one item per 750 ms by changing the interstimulus interval to 500 ms—the experiment was essentially identical to the main Experiment 1 except we only included the Single, Alt, Single-Single, and Alt-Alt conditions. The same pattern was found: There was a main effect of Serial Position, \( F(7, 175) = 89.07, MSE = 0.02, p < .001 \); a main effect of List-Type, \( F(1, 25) = 19.90, MSE = 0.03, p < .001 \); no main effect of Lead-In, \( F < 1 \); but again a significant interaction between Lead-In and List-Type, \( F(1, 25) = 9.15, MSE = 0.02, p < .01 \), whereby the TVE was larger with a lead-in. The results indicate that using a relatively fast presentation rate to investigate the functional characteristics of the TVE is unlikely to compromise the generalizability of the results.
examined these alternative interpretations of the data of Experiment 1.

**Experiment 2**

We sought to rule out the PI/distraction accounts and sought further evidence for the perceptual–gestural mismapping account of the results of Experiment 1 by two convergent means. First, the lead-in on this occasion consisted of letter names rather than digits. This should minimize the likelihood of PI based on the postcategorical similarity of the items used in the lead-in and the TBR digit list (e.g., Bunting, 2006). If such a lead-in still accentuates the TVE, a PI-based explanation seems unlikely. Second, we manipulated the voice conveying this (single-voice) letter lead-in. According to the perceptual–gestural mismapping account, the TVE was augmented by the presence of a lead-in because the voice(s) conveying that lead-in perceptually captured the same-voice items in the alternating TBR list through their similar acoustic (not postcategorical) attributes. Indeed, it has been shown that for auditory stream biasing to occur, the tone(s) in the lead-in must be very similar in frequency to those in an alternating-tone test sequence (Anstis & Saida, 1985). Thus, in Experiment 2, we contrasted a condition involving a single-voice lead-in spoken in one of the voices conveying the alternating TBR list (as in the Single-Alt condition of Experiment 1) and one involving a single-voice lead-in spoken in a third voice—a voice different from either of the voices conveying the alternating voice TBR list. On the basis of the perceptual–gestural mismapping account, this different voice lead-in—because of its relative acoustic dissimilarity to the list voices particularly in terms of frequency/timbre—should be less effective in promoting by-voice partitioning of the alternating-voice TBR list and, hence, in augmenting the TVE. Thus, in relation to Table 2—which shows all six trial types contrasted—whereas a same-voice lead-in (as in Experiment 1) should further impair recall of an alternating voice list as compared with the Alt condition, such accentuation should be diminished with a different-voice lead-in. On the PI or distraction accounts, there is little reason to suppose that the particular voice conveying the lead-in should modulate its impact.

**Method**

**Participants.** Twenty-six undergraduates from Cardiff University, each reporting normal hearing and normal or corrected-to-normal vision, took part in return for course credits.

**Apparatus and materials.** The TBR lists were constructed and presented in the same fashion as Experiment 1. However, in addition to recording the digit sets 1–8, the letters that now served as the lead-in—r, s, t, u, v, x, y, and z—were digitally recorded three times, once in the same female voice and once in the same male voice used for the digit sets (the same voices as used in Experiment 1) and once in a different (male) voice chosen on the basis that it was clearly distinguishable perceptually from both the other two (male and female) voices. Table 2 shows the full set of six conditions: All TBR lists were presented in one of the original voices or in both original voices presented in an alternating fashion. In the conditions involving a lead-in, a letter countdown (r through z excluding w) was presented either in the same voice as that conveying an ensuing single voice list (Same-Voice/Single) or in the different voice (Different-Voice/Single). For the Same-Voice/Alt condition, the lead-in was presented in the same voice as that conveying the second, fourth, sixth, and eighth items of the ensuing alternating-voice list. Finally, in the Different-Voice/Alt condition, the lead-in was presented in the different voice and, hence, was distinct from both voices conveying the ensuing alternating-voice list.

**Design and procedure.** The design and procedure were the same as Experiment 1 except that the three levels of the Lead-In factor were no lead-in, same-voice lead-in, and different-voice lead-in. The 56 trials in the with lead-ins block were composed of 14 Same-Voice/Alt trials, 14 Different-Voice/Alt trials, 14 Same-Voice/Single trials, and 14 Different-Voice/Single trials and was preceded by eight practice trials: two from each of the four conditions. The without lead-ins block comprised 28 trials made up of 14 single-voice lists and 14 alternating-voice lists preceded by four practice trials: two from each condition. For further counterbalancing measures, see the Appendix. The procedure was the same as for Experiment 1.

**Table 2**

Schematic Representation of the Six Conditions Contrasted in Experiment 2

<table>
<thead>
<tr>
<th>Condition</th>
<th>Voice</th>
<th>Lead-in</th>
<th>To-be-remembered list</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Single</td>
<td>Female (or male) voice:</td>
<td>6 5 2 7 1 4 8 3</td>
<td></td>
</tr>
<tr>
<td>2. Alt</td>
<td>Male (or female) voice:</td>
<td>6 2 1 8 5 7 4 3</td>
<td></td>
</tr>
<tr>
<td>3. Same-Voice/Single</td>
<td>Female (or male) voice:</td>
<td>R S T U V X Y Z 6 5 2 7 1 4 8 3</td>
<td></td>
</tr>
<tr>
<td>4. Same-Voice/Alt</td>
<td>Male (or female) voice:</td>
<td>R S T U V X Y Z 6 2 1 8 5 7 4 3</td>
<td></td>
</tr>
<tr>
<td>5. Different-Voice/Single</td>
<td>Different (male) voice:</td>
<td>R S T U V X Y Z 6 5 2 7 1 4 8 3</td>
<td></td>
</tr>
<tr>
<td>6. Difference-Voice/Alt</td>
<td>Male (or female) voice:</td>
<td>R S T U V X Y Z 6 2 1 8 5 7 4 3</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* For Conditions 3–6, the first part of each condition label refers to the voice-presentation format of the lead-in, whereas the second part refers to that for the to-be-remembered list. Single = single voice; Alt = alternating voices.
Results and Discussion

It is clear from Figure 2 that performance was markedly im-
paired in all talker-variable conditions compared with the Single-
Voice conditions (both with and without a lead-in). Moreover,
whereas performance was worse in the Same-Voice/Alt condition
compared with the Alt condition, this was not the case for the
Different-Voice/Alt condition. Thus, in line with the perceptual-
gestural account, the TVE was accentuated by a lead-in only when
that lead-in was in the same voice as one of the voices conveying
the ensuing alternating-voice sequence (Same-Voice/Alt).

A 2 (List-Type) × 3 (Lead-In) × 8 (Serial Position) repeated
measures ANOVA revealed a main effect of Serial Position, $F(7,
189) = 210.27$, $MSE = 0.038$, $p < .001$; and of List-Type, $F(1,
27) = 111.42$, $MSE = 0.040$, $p < .001$; but no main effect of
Lead-In, $F(2, 54) = 2.82$, $MSE = 0.024$, $p = .068$. Importantly,
the List-Type × Lead-In interaction was significant, $F(2, 54) =
3.51$, $MSE = 0.018$, $p < .05$. Simple effects analyses showed that
whereas there was no significant difference between performance
in the Different-Voice/Alt condition and the Alt condition ($p >
.05$), performance was indeed poorer in the Same-Voice-Alt con-
dition compared with the Alt condition ($p < .001$). Again, there
was no difference between any of the conditions with a single-
voice list (Single, Same-Voice/Same, Different-Voice/Single),
showing that the lead-in in-and-of itself did not affect perform-
ance. Both Lead-In and List-Type interacted with Serial Position
($p < .05$); however, we refrain from attempting to ascribe any
theoretical significance to these interactions.

The results of Experiment 2 indicate that the lead-in’s power to
augment the TVE in Experiment 1 was not dependent on the fact
that it comprised items that were postcategorically similar (in fact
identical) to those to be recalled: A letter lead-in conveyed in the
same voice as one of the voices in the ensuing TBR list also
accentuates the TVE. Moreover, a letter lead-in has no effect if the
voice in which it is conveyed is relatively acoustically dissimilar to
both voices in the TBR list. This is entirely as would be expected
on the basis of previous evidence that auditory stream biasing is
highly sensitive to the similarity of the frequency of lead-in and
test tones (Anstis & Saida, 1985). Together, these two aspects of
the results of Experiment 2 provide strong support for the view that
a (same-voice) lead-in has its impact not by being postcategorici-
cally similar to the TBR list but by capturing the same voice items
in the following TBR list through their similar acoustic character-
istics. Neither a PI-based (e.g., Bunting, 2006) nor a general
attentional distraction-based explanation can readily account for
this pattern of results.

The results of Experiments 1 and 2 are also not readily ex-
plained by previous item-decay accounts of the TVE (Goldinger et
al., 1991; Martin et al., 1989): It is far from clear on this approach
why the presence of a lead-in generally should accentuate the TVE
and, more particularly, why that lead-in has to be acoustically
similar to one of the voices in the alternating list to exert this
effect. Indeed, one might have expected preexposure to the indexi-
cal properties (e.g., timbre) of the same voice as that conveying
some of the ensuing TBR list (Single-Alt condition in Experiment
1 and Same-Voice Lead-In/Alt in Experiment 2) or both of the
voices conveying that list (Alt-Alt condition of Experiment 1)—
and particularly being preexposed to the temporal pattern of voice
changes (Alt-Alt condition)—to facilitate voice normalization or
incorporation processes. Such facilitation should in turn have
allowed greater opportunity to refresh decay-prone item-traces via
rehearsal and hence reduce, rather than accentuate, the TVE.

Experiment 3

In Experiment 3, we take a different approach to testing the
perceptual–gestural mismapping account. The account yields the
perhaps counterintuitive prediction that increasing the number of
different voices in a TBR list should diminish, not augment, the
TVE. This follows on the grounds that the smaller the number of
different voices, the more likely and often it is that nonadjacent
items will be more acoustically similar to one another and hence
the greater the likelihood of their forming a coherent perceptual
stream. Before explaining further, we describe the three conditions
contrasted in Experiment 3 in more detail. The TBR list in the
four-voice condition comprised the female samples used in Exper-
iment 1 and another three voices generated by pitch-shifting those
female (F) samples down by three semitones (hereafter: F↓), up
by three semitones (F↑), and up by six semitones (F↑↑). The
eight-digit list was conveyed in the following pattern of voices: F
F+ F↓ F↓ F↓ F↓ F↓ F↓ (or its mirror image: F+ F F F→ F F+ F↓)
F↓ F↓). In the single-voice condition, the TBR list was
conveyed in either one of the four voices, whereas in the
alternating-voice condition, the TBR list involved an alternation
between F and F+ (or F+ and F). For each list-type, given that

![Figure 2](image-url)

Figure 2. Mean percentage of items correctly recalled at each serial
position in the six conditions of Experiment 2 (see Table 2 for an illus-
tration of the six conditions). Single = single voice; Alt = alternating voices;
Diff = different.

4 We purposefully pitch-shifted the original female voice rather than
using the original sets of male and female voices and recording another two
additional talkers so that the degree of acoustic difference between each
successive pair of adjacent items was roughly equal for the two- and
four-voice conditions (this was also the reason for choosing the particular
pattern of voice changes used in the four-voice condition). If we had used
recordings from four different talkers (or used a different pattern), it would
be very difficult to manipulate “number of different voices” without
confounding this factor with differences in the degree of acoustic differ-
ence between successive items.
we have shown that the TVE is more robust when lead-ins are used, each TBR list was preceded by a lead-in (a digit countdown as in Experiment 1) in which the voice or pattern of voices conformed to that characterizing the ensuing TBR list.

On the basis of the perceptual–gestural mismapping account, we expect poorer recall in both the alternating- and four-voice conditions than in the single-voice condition because of the far greater perceptual incoherence of successive items in the two talker-variable conditions. However, the potentially more telling prediction is that the alternating-voice list should produce the poorest performance. This is because it is well known that the likelihood of temporally nonadjacent items perceptually “capturing” one another into the same stream is a function of both their acoustic similarity, the number of times those similar items are encountered, and the temporal proximity of the acoustically similar events (Bregman, 1990). Thus, given that in the alternating-voice list, there are six instances in which immediately nonadjacent items are in the same voice, whereas there are only two such instances in the four-voice list, the propensity for nonadjacent items to perceptually capture one another to form a coherent stream—and hence the degree of perceptual–gestural mismapping—is greater in the alternating- compared with the four-voice list.

Method

Participants. Forty undergraduates from Cardiff University took part in return for course credits. Each participant reported normal hearing and normal or corrected-to-normal vision.

Apparatus, materials, design, and procedure. These aspects of the method were the same as Experiment 1 except for the following: Three new sets of voice samples were generated by pitch-shifting the original female-spoken items down by three semitones, up by three semitones, and up by six semitones (without altering each item’s duration) using the “pitch-shift” function in the Soundforge 7 software. The two repeated measures factors were List-Type (three levels: single-voice, alternating-voice, and four-voice) and Serial Position (eight levels). There was one block with 84 trials: 28 single-voice trials, 28 alternating-voice trials in which the F voice alternated with the F+ voice, and 28 four-voice trials (forming either the pattern F F+ F++ F+ F F− F F+ or F+ F F− F F+ F++ F+ F). Each TBR list was preceded by a lead-in that conformed to the same voice-format as that TBR list. One of each of the eight trial types was presented as practice trials.

Results and Discussion

It is evident from Figure 3 that performance was markedly impaired in both talker-variable conditions compared with the single-voice condition. More importantly, in line with the perceptual–gestural mismapping account, performance was significantly poorer in the alternating-voice condition than in the four-voice condition. There was a main effect of Serial Position, F(7, 273) = 120.72, MSE = 0.03, p < .001; and of List-Type, F(2, 78) = 49.87, MSE = 0.02, p < .001; as well as an interaction between List-Type and Serial Position, F(14, 546) = 5.57, MSE = 0.006, p < .001. Planned repeated contrasts showed that performance in the four-voice condition was significantly poorer than in the single-voice condition, F(1, 39) = 63.27, MSE = 0.05, p < .001, and, most importantly, that performance in the alternating-voice condition was slightly (Cohen’s d = 0.12) but significantly poorer than in the four-voice condition, F(1, 39) = 5.9, MSE = 0.02, p = .02.

The results of Experiment 3 provide further support for the perceptual–gestural mismapping account: Whereas the perception of true temporal order would be impaired by the lack of perceptual coherence in both talker-variable conditions, the nonadjacent items in the alternating-voice list condition would be more strongly grouped on account of their greater acoustic similarity and the greater number of repetitions of those more similar events (voices) compared with the case in the four-voice condition (see, e.g., Bregman & Rudnicky, 1975). Again, it is unclear how the item-decay accounts could accommodate this result: There is no apparent reason why voice normalization (Martin et al., 1989) or incorporation processes (Goldinger et al., 1991) would be slower—hence leading to more item decay—with two different voices in the list as compared with four.

The present results also militate against a suggestion made by Greene (1991) that the TVE may result from participants adopting a deliberate, but counterproductive, strategy of grouping (or rehearsing) the TBR items by voice when presented with an alternating-voice list. This suggestion draws on evidence from free recall paradigms in which participants tend to cluster at output items that share source characteristics such as language (Tulving & Colotla, 1970), modality (Murdoch & Walker, 1969), or type-face (Hintzman, Block, & Inskip, 1972; for a recent review, see Polyn, Norman, & Kahana, 2009). On this deliberate-grouping view, the TVE reflects the cost of having to reorganize the items into their true serial order at output having deliberately organized the items by voice during presentation. This differs from our account in which the nonstrategic, obligatory grouping of items by voice conflicts with the deliberate attempt to assemble the items according to their true temporal order; at no point is there a deliberate attempt to organize the items by voice. Given that it seems unlikely that participants would choose to adopt a deliberate strategy of grouping the items by voice in the four-voice condition (i.e., into
four groups comprising two items each), the deliberate grouping account has difficulty in explaining the marked impairment in this condition compared with the single-voice condition. Even if a four-group strategy (in the four-voice condition) is feasible, it is reasonable to expect that it would be more burdensome to reorganize the items into canonical order at output than with a two-group strategy (in the two-voice condition), an expectation at odds with the present results.

The findings of Experiments 1–3 converge to support the view that one key component of the TVE is an obligatory auditory perceptual organization of the items that conflicts with their true temporal order. The second key aspect of our account is that the locus of the impairment is, ultimately, the gestural sequence-planning process: The process of assembling the items into a gestural analogue is fed impoverished or incompatible information regarding the order of the items by the obligatory auditory perceptual organization process. In Experiment 4, we turn to focus on this gestural-planning component of the account.

Experiment 4

The analytical device of examining whether two or more variables known to independently affect serial recall combine to produce an additive or nonadditive effect has played a key role in the development of short-term memory theory (e.g., Baddeley, Lewis, & Vallar, 1984; Larsen & Baddeley, 2003; Longoni, Richardson, & Aiello, 1993). In Experiment 4, we examine for the first time the possible interplay of talker variability and the phonological similarity effect, a benchmark finding in serial recall whereby phonologically similar items (b, d, v . . .) are more difficult to serially recall than phonologically dissimilar items (f, r, q . . .) (Baddeley, 1966). Within the phonological loop model—the most successful instantiation of the decay-rehearsal based approach—the phonological similarity effect is the chief empirical signature of the passive phonological store (e.g., Baddeley, 2007). However, recent evidence suggests instead that the locus of the effect is the speech-planning process, not a separate passive phonological store: The effect disappears when rehearsal is blocked by articulatory suppression for both visual and auditory lists (Jones et al., 2006, 2004). The residual phonological similarity effect found for auditory lists under suppression—previously taken as evidence against identifying the phonological similarity effect with the rehearsal process (e.g., Baddeley et al., 1984)—is better explained by recourse to acoustic-based perceptual organization processes rather than a phonological store (Jones et al., 2007, 2004; but see Baddeley & Larsen, 2007).

The different accounts of the phonological similarity effect held by the decay-rehearsal model (e.g., Baddeley, 2007) and the perceptual–gestural framework (e.g., Jones et al., 2004) provide a further means of adjudicating between the item-decay accounts and the perceptual–gestural mismatching account of the TVE. A key finding cited in support of the fractionation of the phonological loop into a passive phonological store and an articulatory rehearsal process is that the phonological similarity effect combines additively with the word-length effect: “. . . articulatory duration and the phonemic confusability of items to be remembered exert additive and independent effects upon performance in immediate serial recall, and hence [this shows] that they reflect distinct components of the working-memory system” (Longoni et al., 1993, p. 14; see also Baddeley et al., 1984). That is, whereas the phonological similarity effect reflects a confusion during retrieval between similar item traces in the passive store, the word-length effect reflects a race between articulatory rehearsal and item decay. Thus, given that the item-decay accounts of the TVE appeal to the same mechanism as the decay-rehearsal model offers for the word-length effect, they also predict that phonological similarity and talker variability should exert independent (i.e., additive) effects. In contrast, the perceptual–gestural view posits that both phenomena affect the speech-planning process and hence predicts that the two effects will interact (i.e., will be nonadditive): Phonological similarity will have less impact because the speech-planning process is already impaired by talker variability. In Experiment 4, therefore, participants serially recalled six letters that could either be phonologically similar or dissimilar, and these two types of lists were presented either in a single voice or in alternating voices.

Method

Participants. Twenty undergraduates from Cardiff University took part in return for course credits. Each participant reported normal hearing and normal or corrected-to-normal vision.

Apparatus and materials. Four list types were generated. Each list comprised six letters that were either phonologically dissimilar (k, q, h, v, r, m) or similar (p, d, l, v, b, g), and these two list-types could be presented either in a single voice or in alternating voices. Female- and male-spoken versions of the items were recorded, edited, and presented in the same manner as in Experiment 1.

Design and procedure. There were three repeated measures factors: Serial Position, Phonological Similarity (similar vs. dissimilar), and Voices (single vs. alternating). There were two blocks of trials: 20 phonologically dissimilar lists and 20 phonologically similar lists. Block order was counterbalanced across participants. Within each block, 10 lists were presented in a single voice (five female, five male) and 10 in alternating female–male voices (five starting with a female-spoken item, five starting with a male-spoken item). No trial type was presented more than twice in succession within a block. (Note that none of the lists were preceded by a lead-in in this experiment.) The procedure was the same as Experiment 1 except that before each block, the six consonants to be used in that block were presented in a circle on the screen for 2 min to allow the participants to familiarize themselves with the closed item set (cf. Henson, Norris, Page, & Baddeley, 1996). There were four practice trials (one from each condition) before the experimental trials. The experiment took approximately 25 min.

Results and Discussion

It is apparent from Figure 4 that the phonological similarity effect is diminished in the context of alternating-voice lists. Thus, the data exhibit a nonadditivity of phonological similarity and talker variability in line with the perceptual–gestural mismatching account but at variance with an item-decay approach. A repeated measures ANOVA showed a main effect of Serial Position, $F(5, 95) = 83.39, MSE = 0.03, p < .001$; a main effect of Phonological Similarity, $F(1, 19) = 68.51, MSE = 0.11, p < .001$; a main effect of Voice, $F(1, 19) = 16.15, MSE = 0.04, p < .001$; and, most
importantly, a significant interaction between Phonological Similarity and Voice, $F(1, 19) = 13.16$, $MSE = 0.03$, $p < .01$. The interactions between Serial position and each of the other two variables were also significant, which were subsumed within a significant three-way interaction between all three variables, $F(5, 95) = 3.72$, $MSE = 0.01$, $p < .01$. Again, we do not ascribe any theoretical significance to these interactions at this point given that neither of the accounts being contrasted here makes clear predictions with regard to serial position.

Experiment 4 showed that talker variability and phonological similarity interact (i.e., are nonadditive), consistent with the view that the two effects share a functional locus. The results are therefore at odds with the predictions of item-decay accounts that are based on the decay-rehearsal model, at least as exemplified by the phonological loop model (Baddeley, 1986): If the phonological similarity effect is the empirical signature of a passive phonological store, there is no reason to expect it to interact with the TVE, which from this perspective, has been attributed to the articulatory control process (as with the word-length effect; e.g., Martin et al., 1989). In contrast, the nonadditivity of talker variability and phonological similarity effects is consistent with the perceptual–gestural account: If, as we suppose, talker variability impairs the sequence output-planning process, the expression of another effect also associated with that process (Jones et al., 2006, 2004; Page, Madge, Cumming, & Norris, 2007) would be expected to be diminished with talker-variable lists. Further evidence that the by-voice perceptual organization of an alternating-voice list—as demonstrated in Experiments 1–3—has its impact by conflicting with the assembly or/and rehearsal of a sequence-output plan comes from studies manipulating the nature of the memory task: Tasks that call for the retention of order—and hence promote the use of an articulatory sequence-planning strategy—are far more susceptible to a TVE than those that only require the retention of item information (Hughes, Marsh, & Jones, 2009).

During the peer-review process, it was suggested that the perceptual–gestural mismapping account might also predict an increase in a particular type of order error, whereby nonadjacent items are transposed more frequently in an alternating-voice list than in a single-voice list (whereas adjacent transpositions are by far the most usual type of order error; see, e.g., Henson et al., 1996). However, this prediction does not necessarily follow: With an alternating-voice list, it is possible that it is a tendency to mistakenly recall (rather than transpose) the items by voice that is increased. This would in itself be an error given that recall is scored according to strict temporal order. Importantly, such by-voice recall could potentially result in an increase in the most common, that is, immediately-adjacent, transposition error. Consider the list ABCDEFGH. If alternating voices tend to promote by-voice output, then the list might be recalled as AB-DC-E-GF-H (by-voice output indicated by the dashes). Thus, recalling by voice would have led to adjacent transposition errors between C and D and between G and F in this case. Thus, although it is plausible that participants might also show an increase in nonadjacent transpositions (because of transposing by voice), whether this would be expected to be greater than the increase in adjacent transpositions (because of recalls by voice) is unclear. In the event, transposition analyses of the data from Experiments 1–4—not reported fully here for the sake of brevity—showed that although alternating-voice lists increased the overall number of transpositions, there was no consistent evidence that such lists disproportionately increased a particular type of (e.g., nonadjacent) transposition error.

**General Discussion**

The impact of the present series may be summarized as follows: Experiments 1 and 2 showed that promoting the perceptual partitioning of talker-variable lists by voice by presenting a (same, but not different, voice) single- or alternating-voice lead-in before an alternating-voice list accentuates the TVE in serial recall. Experiment 3 showed that the recall of a two-voice (i.e., alternating) list was poorer than that of a four-voice list, again implicating the action of perceptual streaming processes. Experiment 4 showed that the phonological similarity effect is far weaker in the context of talker-variable lists, which is in line with the perceptual–gestural view that both effects are located in the gestural–sequencing process rather than reflecting distinct components of a bespoke memory system.

**Implications for the Decay-Rehearsal Approach to Serial Short-Term Memory**

Previous accounts of the TVE have been set within the framework of the classical decay-rehearsal approach to verbal short-term memory (e.g., Baddeley, 1986). The process of normalizing out or incorporating indexical attributes (e.g., accent, emotional state) of spoken words delays the encoding or/and rehearsal of each item, thereby promoting item-decay from a phonological short-term store (Goldinger et al., 1991; Martin et al., 1989). The present results are not easily accommodated within this approach. One potential objection, however, is that the item-decay accounts were developed in the context of a “nonpure” serial recall task in which, unlike the present experiments, a new set of items was presented on each trial (e.g., Goldinger et al., 1991; Martin et al., 1989). However, there is little reason why the accounts should not apply in principle to the TVE in pure serial recall: On the decay-rehearsal model, the same item-decay approach is used to explain the word-length effect found in pure serial recall (e.g., Baddeley et al.,...
Nevertheless, the present results do not rule out the possibility that item-decay plays a role in the TVE in nonpure serial recall in which there is a large burden on item as well as order retention, unlike the task used in the present series. An alternative possibility consistent with the perceptual–gestural mismatching account is that although a nonpure serial recall task places a large burden on item memory, talker variability nevertheless impairs only that component of the task that taps the efficacy of a sequence output-planning process. In line with this analysis, in nonpure serial recall (e.g., Goldinger et al., 1991) and also under free recall instructions (Watkins & Watkins, 1980), the TVE is confined to that part of the list—the early part—that has been shown (even in free recall) to be supported by serial rehearsal (Beaman & Jones, 1998; Bhatarah, Ward, & Tan, 2008; Kahana, 1996). Clearly, further research that contrasts the TVEs found in pure and nonpure serial recall will be needed to determine more conclusively whether they share a common mechanism.

It is also important to recognize that TVEs have been observed in many other tasks that seem not to require order retention at all, such as long-term word recognition (Mulhennix, Pisoni, & Martin, 1989), vowel perception (Verbrugge, Strange, Shankweiler, & Edman, 1976), and the speeded classification of words (Mulhennix & Pisoni, 1990). Such effects suggest strongly that the indexical and acoustic features of spoken words are not normalized out during perception but incorporated as part of the encoding episode (e.g., Goldinger, 1996). The present findings lend further support to this general view; if list items are represented in an abstract–phonological form (e.g., Baddeley, 2007), it is difficult to account for the streaming–by–voice effects observed here. However, our findings suggest that this voice incorporation process does not cause the TVE in (at least pure) serial recall by promoting item–decay but may indeed do so through its influence on supra–item perceptual organization.

Although current item–decay based accounts do not fare well in relation to the present findings, there may be other means by which the decay–rehearsal model within which they are conceptually embedded (e.g., the phonological loop model; Baddeley, 1986, 2007) could potentially account for the TVE. One possibility might be to appeal to computational models of how the phonological loop represents serial order, a consideration that has typically been seen as complementary, but secondary, to the core item–based architecture of the underlying functional–level theory (Baddeley, 2007). Thus, talker variability might be seen as disrupting the mechanism supporting serial order in the phonological loop (e.g., a primacy gradient, Page & Norris, 1998; an oscillator–based timing signal, Burgess & Hitch, 1999). However, the interaction between phonological similarity and talker variability in Experiment 4 still seems to pose difficulties for an approach that appeals to an order–based account of the TVE but nevertheless adheres to the concept of a passive phonological store. In computational models of the phonological loop, the phonological content (and hence phonological similarity) of the items only has an impact at a second item–retrieval stage that is independent of whatever mechanism is responsible for representing order (e.g., Burgess & Hitch, 1999; Page & Norris, 1998). Thus, given that the phonological similarity effect is identified with the second item–retrieval stage on this account, its magnitude should not be influenced by a factor—talker variability—presumed to affect a separate order–storing stage, contrary to our data. Moreover, such models rarely consider prephonological perceptual organization processes in explanations of serial recall performance (but see Page & Norris, 1998). Yet, the present results add to a growing body of evidence suggesting that perceptual organization coupled with output–planning processes may alone account for verbal serial recall phenomena without the encumbrance of a bespoke phonological store (e.g., Jones et al., 2006, 2004; Woodward et al., 2008; but see Baddeley & Larsen, 2007; Page et al., 2007).

Implications for Item–Interference Accounts of Serial Short–Term Memory

We have focused on the decay–rehearsal model (e.g., Baddeley, 1986, 2007) of short–term memory as the main theoretical counterpoint to our perceptual–gestural account (but see Experiment 2). However, the TVE seems problematic also for accounts of serial short–term memory in which the relative distinctiveness of items—and hence their relative immunity from interference—serves as the primary explanatory mechanism (e.g., Brown, Neath, & Chater, 2007; Lewandowsky et al., 2008; Nairne, 1990). Although these accounts, like the perceptual–gestural view, eschew a dedicated short–term store, they still invoke a specifically mnemonic process, namely, interference between similar memory traces. For example, according to the feature model (Nairne, 1990; Neath, 2000), serial recall performance is assumed to bear a simple positive relationship to the distinctiveness—and hence immunity from being overwritten—of the items in a TBR list in terms of both their modality–dependent features (e.g., pitch) and their modality–independent features (those that do not vary with modality of presentation). As pointed out previously by Greene (1991), it follows quite straightforwardly that each item in a talker–variable list should be less prone to being overwritten by its successor because there is less overlap in terms of modality–dependent features (e.g., pitch, timbre) than with a single–voice list. Such models cannot, therefore, readily explain why the greater distinctiveness of items in a talker–variable list leads to much poorer, not better, recall (Greene, 1991). Moreover, even if such models could be modified to accommodate the basic TVE, it is not immediately apparent how they might explain the accentuating effect of a lead–in (Experiments 1 and 2) or the nonmonotonic relationship between the level of impairment and the number of different voices (Experiment 3).

Perceptual–Gestural Account: Toward an Embodied View of Short–Term Memory

The present results are in line with previous studies demonstrating the powerful effects of auditory perceptual organization on order processing generally (e.g., Bregman & Campbell, 1971; Creel, Newport, & Aslin, 2004; Warren et al., 1969) and on serial recall in particular (e.g., Macken et al., 2003; Nicholls & Jones, 2002). In the present case, the obligatory by–voice perceptual organization of talker–variable lists gives rise to a perception of order that conflicts with the process of assembling the items into an articulatory plan designed to increase the transitional probabilities between adjacent items. Our appeal to general–purpose processes involved in perception and action that are co–opted to meet the demands of a short–memory task (for similar views, see Glen–
berg, 1997; Reisberg, Rappaport, & O’Shaughnessy, 1984; Wilson & Fox, 2007) resonates with a current shift in cognitive science toward embodying cognition (e.g., Clark, 2006). This shift has emerged as a reaction to the received view of cognition as the action of static, central, and context-free processing and storage structures/resources that are divorced from the so-called “peripheral” processes of perception and action (e.g., Clark, 2006; Hurley, 2001). Instead, an embodied analysis focuses on the dynamic processes involved in goal-directed and coherent engagement with the environment given the constraints and capacities of the organism’s sensorimotor apparatus.

Couched within the embodied cognition approach, one way of fleshing out the gestural component of our account is to suppose that sequence output planning (or “rehearsal”) reflects the operation of a motor-action emulator. Recent work on motor control suggests that for motor-action to be executed in a fluent manner, a forward model of the action—consisting of both the instructions to the effectors and, importantly, the sensory sequelae of the action—is generated so that the imminent action can be compared with the intended action (e.g., Grush, 2004; Schubotz, 2007). An important feature of these models that aligns with the phenomenology of subvocal rehearsal during serial recall is that they can be run without being implemented, that is, they may be run in emulation mode without necessarily resulting in any overt action (e.g., Jordan & Rumelhart, 1992). Thus, we contend that articulatory rehearsal during verbal serial recall is not in the service of refreshing decay-prone items in a labile phonological store (see also Reisberg et al., 1984); rather, the necessarily sequential nature of the emulation of the movements of the vocal tract endows the individual with an ideal medium for taking a series of largely sequentially unrelated verbal items and placing them onto a common carrier, that is, a single, relatively more sequentially coherent, motor plan. The TVE on this view reflects the impairment of an auditory-motor mapping process (cf. Buchsbaum & D’Esposito, 2008), whereby the process of populating the necessarily abstract motor-output emulator system with specific content is ill-informed by auditory-perceptual organization. Another hypothesis that flows from our account, therefore, is that conditions under which perceptual organization yields information that corresponds particularly well to the way in which items need to be assembled into the motor plan should facilitate rather than impair performance. For example, one might contrast two versions of a pure serial recall task, one in which the recall of a subset of adjacent items is preceded (“recall Items 1–5 or 6–10” of the following 10-item list”) and one in which a subset of nonadjacent items is preceded (“recall Items 2, 4, 6, 8, and 10 [or 1, 3, 5, 7, and 9]”; cf. Penney & Butt, 1986). The perceptual–gestural account predicts that alternating-voice lists would impair the first version of the task but facilitate the second. The account might also provide a way of reconstruing the impairment of serial recall when successive items or pairs of items are presented to different sensory modalities (Penney & Butt, 1986) or to different ears (Moray, 1960; Treisman, 1971). For example, ear alternation effects, like the TVE, have typically been attributed to a delay in the processing of each item. The “attention shifting time would reduce the time available for perception and storage” (Treisman, 1971, p. 164). However, as suggested previously by ten Hoopen (1996), these effects may instead be driven by streaming-by-location, again giving rise to a perceptual–gestural mismatching. The device of introducing (on this occasion ear-alternating) lead-ins may provide one means of testing this possibility.

Encouragingly, conclusions from several research programs are now converging on an embodied conceptualization of short-term memory (Acheson & MacDonald, 2009; Buchsbaum & D’Esposito, 2008; Jones et al., 2006, 2004; Postle, 2006; Wilson & Fox, 2007). For example, Wilson and Fox (2007) recently found that serial recall of novel sequences of hand gestures exhibits several of the effects that are, putatively, hallmarks of a specifically verbal short-term memory system, namely, the “phonological” similarity effect, the “articulatory” suppression effect, and the “word” length effect. The authors concluded that “rather than involving hard-wired and dedicated components, working memory may instead consist of the strategic recruitment of cognitive resources, determined on the fly by the immediate demands of the task” (p. 473). We suggest that the present findings illustrate further the fruitfulness of this shift in research focus away from delineating the properties of bespoke stores and mechanisms and toward examining instead how perceptual and action-planning processes support and constrain the retention and reproduction of serial order over the short term.

References
PERCEPTUAL–GESTURAL (MIS)MAPPING IN SERIAL RECALL


Maybery, M. T., Parmentier, F. B. R., & Jones, D. M. (2002). Grouping of


Appendix

Additional Counterbalancing Measures Incorporated Into the Design of Experiments 1–3

Experiment 1

The 14 trials within each condition were further subdivided as follows (note that Single = single voice and Alt = alternating voices):

With Lead-In Block

- 14 Alt-Alt: Seven trials in which the to-be-remembered (TBR) list started with a female item and seven in which it started with a male item;
- 14 Alt-Single trials: Seven in which the TBR list was female spoken and seven in which it was male spoken;
- 14 Single-Single trials: Seven in which the TBR list was female spoken and seven in which it was male spoken;
- 14 Single-Alt trials: Seven in which the TBR list started with a female item and seven in which it started with a male item.

Without Lead-In Block

- 14 Single TBR lists: Seven female lists, seven male lists;
- 14 Alt TBR lists: Seven started with a female item first, seven started with a male item first.

Experiment 2

The 14 trials within each condition were further subdivided as follows:

With Lead-In Block

- 14 Same-Voice/Alt trials: Seven in which the lead-in was presented in a female voice and the TBR list started with a female item and seven in which the lead-in was presented in a male voice and the TBR list started with a male item;
- 14 Different-Voice/Alt trials: Seven in which the TBR list started with a female item and seven in which it started with a male item;
- 14 Same-Voice/Single trials: Seven in which both the lead-in and the TBR list were female spoken and seven in which both the lead-in and the TBR list were male spoken;
- 14 Different-Voice/Single trials: Seven in which the TBR list was female spoken and seven in which it was male spoken.

Experiment 3

The 28 trials in each condition were further subdivided as follows:

- 28 Single trials: Seven in each voice: F, F−, F+, and F++;
- 28 Alt trials: 14 started with the F voice, and 14 started with the F+ voice;
- 28 four-voice trials: 14 forming the pattern “F F+ F++ F+ F F− F F+” and 14 forming the pattern “F+ F F− F F+ F++ F+ F.”

Received August 14, 2008
Revision received May 27, 2009
Accepted June 2, 2009