Tweaking the lexicon: Organization of vowel sequences into words

RICHARD M. WARREN, JAMES A. BASHFORD, JR., and DANIEL A. GARDNER
University of Wisconsin-Milwaukee, Milwaukee, Wisconsin

The ability of listeners to distinguish between different arrangements of the same three vowels was investigated for repeating sequences having item durations ranging from 10 msec (single glottal pulses) up to several seconds/vowel. Discrimination was accomplished with ease by untrained subjects at all item durations. From 30 through 100 msec/vowel, an especially interesting phenomenon was encountered: The sequences of steady-state vowels were organized into words, with different words heard for the different arrangements of items. In a second experiment, repeating sequences of random arrangements of 10 40-msec vowels were employed. When sets of four such sequences were presented to listeners, distinctive words were heard, which permitted each arrangement to be discriminated from the others. In addition, minimal differences (reversing the order of a single contiguous pair of vowels) in the 10-item sequences could be detected via verbal mediation. Hypotheses are offered concerning mechanisms responsible for these results.

A succession of steady-state vowels presented loudly and clearly can be heard as a word. This unusual verbal organization can help us understand how acoustic components are processed in speech perception.

In the experiments reported here, repeated or recycled sequences were employed. These iterated stimuli were first used in the 1960s as a means of allowing a limited number of sounds (usually 3 or 4) to be presented for extended periods (Warren, 1968; Warren, Obusek, Farmer, & Warren, 1969). Repetition also helps eliminate the special cues to the order of items that are provided by the first and last items of a sequence (Divenyi & Hirsh, 1978; Warren, 1972). In studies with recycled vowel sequences, it has been shown that the naming or identification of order is accomplished readily at 200 msec per item, but that it is not possible at item durations below 100 msec (Cole & Scott, 1973; Cullinan, Erdos, Schaefer, & Tekieli, 1977; Dorman, Cutting, & Raphael, 1975; Thomas, Cetti, & Chase, 1971; Thomas, Hill, Carroll, & Garcia, 1970; Warren, 1968; Warren et al., 1969; Warren & Warren, 1970). In none of these studies have observations involving vowel durations below the threshold for identification of order been reported. However, in preliminary observations, we found that when three steady-state vowels (A, B, and C) were presented as recycled sequences, the two possible arrangements (...ABCABCA... and ...ACBACBA...) could be discriminated readily at item durations much briefer than the limit for naming of order.

Our first experiment confirmed that discrimination between different orders of the same speech sounds does

not require the ability to identify the order of the phonemes, or indeed even the ability to identify the components within the sequences. Listeners were required to judge whether alternately presented recycled sequences of three vowels (which could be presented in identical or permuted item orders) were *same* or *different*. The vowels spanned the range from 10 msec (single glottal pulses) through 5 sec (500 glottal pulses), with no acoustic mixing or transitional stages in going from one vowel to the next. When vowel durations were above 100 msec, listeners could name the phonemes in the proper order, and they could then use the difference in named order to distinguish the two arrangements. When vowel durations were below 30 msec, resemblance to speech was absent, and differences in quality or timbre made it easy to discriminate between the two arrangements (for example, a listener might report that one sequence sounded "rougher" than the other). Between these values (30-100 msec), listeners could hear different words (usually lexical, sometimes nonsense) for each of the arrangements. The words heard differed across individuals, and they normally bore little resemblance to the actual phonemes.

In other studies too, verbal organization of repeated sequences of vowels has been observed. Dorman et al. (1975) used these stimuli to investigate the limits for identification of temporal order and noted in passing that verbal organizations interfered with the listeners' task when vowel durations approached the lower limit for order identification. Skinner (1936) used repeated sequences of barely audible vowels with durations of several hundred milliseconds. He reported that his listeners heard words and sentences. When the levels of vowels were raised well above threshold (as were the vowels in each of our experiments), "imitative" responses occurred, and the sequences were identified as a succession of vowels (in keep-

This study was supported in part by grants awarded to R. M. Warren from the National Institutes of Health (DC00208) and the Air Force Office of Scientific Research (88-0320). We thank Bradley S. Brubaker for his valuable help. Correspondence may be sent to Richard M. Warren, Department of Psychology, University of Wisconsin-Milwaukee, Milwaukee, WI 53201.

ing with observations made in our Experiment 1 for vowels with durations of a few hundred milliseconds). Skinner attributed the verbal organization of his faint syllabic-length vowels to a "summation" of originally subliminal responses.

In our second experiment, we employed 48 recycled sequences, each consisting of a different random arrangement of a set of 10 steady-state 40-msec vowels played loudly and clearly. Individual listeners heard characteristic words or pseudowords corresponding to each of the orders, and they could identify a particular sequence among several on the basis of its verbal correlate. Interestingly, listeners often heard a particular arrangement as two concurrent words that differed in timbre and/or pitch. As we shall see, this splitting of the stimulus provides a clue to the mechanisms employed for perceptual syntheses.

In another part of the second experiment, we also employed recycled sequences of 10 different 40-msec vowels. The stimuli consisted of pairs of sequences with minimal differences in structure (the orders of two contiguous vowels were interchanged). Listeners were again able to use verbal mediation to distinguish members of the pairs.

EXPERIMENT 1: DISCRIMINATION BETWEEN DIFFERENT ORDERS OF THREE-ITEM VOWEL SEQUENCES

This first experiment was designed in part to test the hypothesis that discrimination between different orders of the same speech sounds does not require the ability to identify the order of the phonemes, or indeed even the ability to identify the components within the sequences. Our listeners were required to judge whether alternately presented recycled sequences of three vowels (which could be presented in identical or permuted item orders) were *same* or *different*. The vowel durations extended from 10 msec to 5 sec, permitting a comparison of discrimination accuracy and cognitive strategies employed for durations corresponding to, as well as briefer and longer than, those occurring in speech.

Method

Subjects. Participants were recruited from introductory psychology courses; they received either course credit or cash for their participation. The students who passed the audiometric screening procedure described below were assigned randomly to one of two experimental groups, each containing 36 subjects.

Audiometric screening. All subjects participating in the experiments passed an audiometric test designed to eliminate individuals with hearing deficits, as well as anyone who failed to follow the standard instructions used with the Békésy threshold tracking procedure. Following presentation of instructions and familiarization with the task, a pure tone presented diotically was swept up from 400 through 9000 Hz, and then down from 9000 through 400 Hz, at a rate of one octave/minute while subjects tracked their thresholds. Tracking was accomplished using a hand-held button switch (depressing the button decreased the intensity at a rate of 2.5 dB/sec, and releasing the button increased the intensity at the same rate). An X-Y plotter produced audiograms consisting of continuous threshold tracings. Subjects were excluded from further participa-

tion if either directional sweep resulted in audiograms that differed from the 1964 ISO standards by more than 22.5 dB at any frequency for the portion of the audiogram extending from 500 through 8000 Hz.

Stimuli. The first stage in the preparation of the recycled sequences of three vowels used as stimuli involved production of extended steady-state recordings of three vowels (/ʌ/, /æ/, /i/) on parallel tracks of a multitrack recorder (16-track Ampex Model MM 1200). These steady-state vowels were derived from recorded statements of syllables containing these vowels ("hud" for /ʌ/, "had" for /æ/, and "heed" for /i/) produced by a male speaker at a vowel fundamental frequency of 120 Hz (the speaker matched the pitch of the vowel to that of a complex tone of 120 Hz heard through headphones). A complete single glottal pulse was excised from the central portion of each consonant-vowel-consonant statement. The waveforms of the glottal pulses were monitored and the period measured by a Nicolet Model 3091 digital storage oscilloscope used in conjunction with a programmable digital delay line (modified Eventide Model BD955) capable of repeating or "looping" stored input corresponding to a single glottal pulse. The repetition period of the delay line was set at 8.33 msec for each of the vowels (which corresponded to a repetition rate of 120 Hz), and recordings of the steady-state vowels were made on parallel tracks.

Two types of series were recorded for each duration employed (10, 12, 30, 100, 300, 1,000, 3,000, and 5,000 msec). A *different* series with successive sequence bursts consisting of /ʌ/, /æ/, /i/, /ʌ/, /æ/, /i/, ... /ʌ/, alternating with the permuted order /i/, /æ/, /ʌ/, /i/, /æ/, ... /ʌ/, and a *same* series with all bursts consisting of /ʌ/, /æ/, /i/, /ʌ/, /æ/, /i/, ... /ʌ/. Note that because of the special ease of identifying the first and last items of a sequence (Warren, 1972), each sequence began and ended with the same item. Table 1 lists the parameters for the stimuli employed, giving item durations, the number of items (vowel statements) in each sequence burst, the interburst interval separating successive bursts (which were either identical or alternating in item order), and the number of bursts constituting a stimulus set.

All sequences (except those with a 12-msec item duration) were generated by gating the output from the three parallel tracks containing extended steady-state single vowels prepared as described above. The output of these tracks was passed through three Coulbourn electronic switches set for a rise/fall time of 2 msec. A series of timers (Grason-Stadler Model 1216A) controlled passage of each vowel through its gate, and introduced a 1-msec separation between the waveforms corresponding to the ending of one vowel and the beginning of the next as seen on the digital storage oscilloscope (this separation minimized the acoustic interaction of items). Another timer regulated the silent interburst interval separating successive bursts. The path of the signals through the equipment was identical in both the *same* and the *different* order series, with the relative timing of the opening and closing of the gates producing

Table 1
Description of the Stimuli Consisting of
Three Recycled Vowels in Experiment 1

Item Duration (in msec)	Items per Sequence Burst	Interburst Interval (in msec)	Bursts per Stimulus Set	Stimulus Set Duration (in sec)
10	301	300	10	32.8
12*	301	300	10	38.8
30	91	300	10	30.0
100	31	300	8	26.9
300	10	300	8	26.1
1,000	10	1,000	4	43.0
3,000	7	1,000	4	87.0
5,000	7	1,000	4	143.0

*Items with locked waveforms.

the permuted orders. The number of vowel statements in the burst was controlled by Coulbourn predetermined counters (Model S43-30). The outputs of the gates were combined in a Yamaha audio mixer (Model PM-430) and band-passed from 100 through 8000 Hz by a Wavetek/Rockland filter (Model 751A) having slopes of 115 dB/octave before recording on one of the tracks of the multitrack recorder reserved for the experimental stimuli. The input level of each of the vowels as delivered to the tape recorder was adjusted separately to produce equal intensity (dBA) for all vowels on subsequent playback through the headphones used by the subjects. Sets of sequences were recorded for each item duration with the track for a *same* set parallel to the track of the corresponding *different* set, so that the experimenter could present either *same* or *different* stimuli at the same tape positions.

In sequences consisting of 10-msec items, only a single statement of each vowel's waveform was gated before switching to the next. Since this programmed switching was not exactly synchronous with the waveform repetition period of the recorded vowels (the recorder had a frequency stability of $\pm 0.1\%$), the repeated sequences underwent slow drifts in their waveforms and perceptual qualities. Sequences with longer item durations consisted of multiple identical statements of each vowel's waveform before switching to the next, and perceptual quality was more stable.

Sequences with 12-msec item durations were constructed with *locked* waveforms, so that switching always occurred at a fixed position in the waveform of each vowel and drifting did not take place. This stimulus was prepared as follows: Three separate delay lines (two modified Model BD955 and one modified Model 1745M Eventides) were driven by a common clock, and each repeated a single glottal pulse of a different vowel (the glottal pulses were derived from the same extended statements of the three vowels used for preparing the other sequences). The manner of capturing and repeating single glottal pulses on the delay lines was similar to that already described, except that the programming equipment introduced a 3.67-msec silent interval between successive statements of glottal pulses. The splice point of each single-vowel digital loop was at the center of this silent interval. The repetition period of all delay lines was set at 12 msec (measured by a common clock), the vowel statements were aligned so that each of the three vowels began and ended synchronously, and the vowels were then recorded simultaneously on separate tracks of the multitrack recorder. A timing signal (a unipolar pulse) generated by one of the delay lines at the splice point of its digital loop was recorded on a fourth track at the same time as the vowels. This recorded timing signal permitted the programming equipment to gate single glottal pulses of each recorded vowel in the desired order. Following gating and mixing, the locked three-item vowel sequences were recorded on an additional channel of the multitrack recorder, as described for the other sequences.

Procedure. The subjects passing the audiometric screening test were recalled for their single experimental session lasting about 40 min. They were tested individually while seated in an audiometric room along with the experimenter. The stimuli were presented diotically through matched headphones at a level of 70 dBA as measured by a sound-level meter with a 6 cc coupler. The experimenter operated the tape recorder (located outside the chamber), using a remote control unit equipped with a preset multi-point rapid search-to-cue device. Switches on an audio mixer permitted delivery of the output from the desired track of the recorder.

Half of the 72 subjects served in the main experiment, which included all the sequence pairs listed in Table 1 except for the sequences with 10-msec vowel durations (that is, item durations of 12, 30, 100, 300, 1,000, 3,000, and 5,000 msec) presented in the order listed. The order of increasing item durations was employed to avoid the possibility (discussed by Warren, 1974) that with a series of decreasing item durations, the naming of orders at brief item durations could be accomplished through recognition of qualita-

tive similarities to the previous sequences having longer item durations with directly identifiable orders. As described earlier, the 12-msec sequences had switching from one vowel to the next vowel locked, so that each restatement of a particular vowel was a single intact glottal pulse. All other sequences were *nonlocked*, with successive statements of each vowel starting and stopping at different waveform positions. The separate group of 36 subjects serving in the supplemental experiment received only the stimulus consisting of 10-msec vowels.

The subjects were told that they would be hearing patterns of sounds separated by brief silent intervals, and that their task was to determine if all patterns were identical or if alternate patterns differed in any way. They were instructed to call out "same" or "different" at any time during the stimulus presentation. They were informed that the occurrence of *same* and *different* groupings would be randomly determined. The subjects were encouraged to ask questions if any part of the instructions was unclear. After both the subject and experimenter were satisfied that the instructions were understood, the sequences were presented in an order of increasing item duration for the 36 subjects in the main experiment (the 36 subjects serving in the supplemental experiment received only the 10-msec items).

Before presentation of unknown *same* or *different* sequences at each item duration, each subject was given sample sequences (first a *different* set, then a *same* set) which were identified by the experimenter as *same* or *different*. They were told that they could hear either of the known samples again, if they wished, before hearing the unknowns. When a subject indicated readiness, he or she was given three unknowns at that item duration. The *same* and *different* unknowns were presented in a pseudorandom order, with the constraint that all three of the unknowns presented to a subject at any item duration could not all be of a single type. No feedback was given concerning the accuracy of judgments with the unknowns. In the main experiment, of the total of 21 unknowns presented to each subject, 10 were the same and 11 were different for 18 subjects, and 11 were the same and 10 were different for the other 18 subjects. Half the subjects received orders of *same* and *different* unknowns that were "mirror images" of the other half, with *same* and *different* unknowns being interchanged to maintain symmetry of unknown groupings. This symmetry was also maintained for the supplemental group receiving the 10-msec vowel durations.

Results

Table 2 shows that the accuracy of discriminating between permuted orders ranged from 78% correct to 99% correct, and that it was significantly better than chance

Table 2
Accuracy of Same/Different Judgments for Recycled Sequences of Three Vowels in Experiment 1

Stimulus Duration	Responses		
	No. Correct (out of 108)	% Correct	Z Scores*
10†	84	78	5.82
12‡	98	91	8.52
30	90	83	6.86
100	98	91	8.52
300	102	94	9.15
1,000	103	95	9.35
3,000	106	98	9.98
5,000	107	99	10.18

Note—Stimulus duration is given in milliseconds for each stimulus item. *All $ps < .0001$. †Judgments made by separate groups. ‡Items with locked waveforms.

for all of the item durations used ($Z \geq 5.77$, $p < .0001$). The sequences consisting of 10-msec items (with slowly drifting waveforms and perceptual qualities) had 78% correct responses, while the 12-msec locked sequences (with switching occurring at fixed points corresponding to the beginning and end of the single glottal pulse representing each vowel) had 91% correct responses. This difference was significant ($Z = 2.62$, $p < .01$).

Questioning of listeners after completion of the formal experiment suggested that two basic ways of discriminating between the different arrangements of items were used: (1) naming of components in their proper order for vowel durations greater than 100 msec; and (2) a holistic recognition of patterns without the ability to identify the order of components (or even the components themselves) for vowel durations from 100 msec down to 10 msec. The range from 100 to 10 msec consisted of two regions: (2a) From 100 to 30 msec, the sequences of three vowels could be heard as words rather than steady-state vowels, with different words heard for the different arrangements; (2b) below 30 msec, the vowel sequences were heard as nonlinguistic sounds, with different qualities associated with the different arrangements. These perceptual categories (1, 2a, 2b) reported by untrained listeners agreed with observations made by laboratory personnel.

Discussion

Limits for the naming of order. The earliest experiments with recycled sequences of sounds measured thresholds for identifying the order of component items (Warren, 1968; Warren et al., 1969; Warren & Warren, 1970). When four 200-msec sounds were used, listeners instructed to name the order of items performed at chance level with unrelated sounds consisting of noises, tones, and buzzes, but they could accurately name the order of vowels. In subsequent studies, it was established that the threshold for identifying the order of unrelated sounds is 300 msec or more (Warren & Obusek, 1972), and that the threshold for correctly ordering the pitches associated with sequences of four sinusoidal tones is between 125 and 200 msec/item (Nickerson & Freeman, 1974; Thomas & Fitzgibbons, 1971; Warren & Byrnes, 1975). The lowest thresholds for four-item sequences (about 100 msec/item) were obtained with vowels (Dorman et al., 1975; Thomas et al., 1971).

There seems to be general agreement that vowel order can be named at briefer durations than is possible with other sounds. Why this difference? Using evidence from several sources, it was proposed by Warren (1974; also suggested independently by Teranishi, 1977) that the time required for verbal labeling or naming of components in extended sequences was the threshold-determining stage in the identification of order. Since vowels have a name that is the same as the sound itself, no recoding is necessary (naming order can be accomplished through a simple echoic restatement of the stimulus items), and the time required for identifying order is minimal. Nevertheless,

as discussed below, the threshold value of 100 msec seems too high for agreement with models that consider identification of phonemic order to be necessary for the comprehension of speech.

Normal conversation has an average duration of speech sounds of about 80–100 msec; this duration drops to about 70 msec for oral reading, and some comprehension of “compressed speech” is possible at average phonetic durations of only 30 msec (for a brief summary of this literature, see Warren, 1982, pp. 119–120). Recognizing that there was a discrepancy between the rate of phoneme occurrence within intelligible speech and the ability to perceive order in a sequence of independently generated speech sounds, Wickelgren (1969) suggested that context-sensitive allophones facilitated temporal ordering. Coarticulation is, at least in part, an acoustic consequence of inertial and neuromuscular constraints on the movement of the tongue and other articulatory organs from one position to the next, and Wickelgren considered that recognition of particular allophonic forms might make it possible to identify more than one phoneme in a single speech sound. Thus, order could be identified at much briefer durations than would be possible for a succession of independent sounds. A number of experiments have demonstrated that coarticulation (and other cues increasing the resemblance of phonetic sequences to normal speech) does indeed facilitate the task of naming components and their orders (Cole & Scott, 1973; Cullinan et al., 1977; Dorman et al., 1975; Warren, 1968; Warren & Warren, 1970). But in no case, even with coarticulation cues, could orders be identified at item durations below 100 msec. However, listeners can comprehend speech consisting of phonemes with average durations of considerably less than 100 msec. One explanation for this discrepancy is that phonetic order is determined at some early level of linguistic processing that is not accessible for the naming of this order. Another hypothesis (which we favor) is that a determination of the order of component speech sounds is not necessary at any level of analysis for the recognition of words or for the comprehension of discourse. It is to be suggested that acoustic sequences need not function as perceptual sequences (that is, a succession of discrete sounds). Patterns formed by particular arrangements of speech sounds may be recognized as *temporal compounds* without any need for identification of constituents. As discussed below, this concept of temporal compound formation was formulated initially on the basis of experiments with nonverbal sounds.

Nonphonetic temporal compounds. In earlier studies involving arbitrarily selected sounds (noises, sinusoidal tones, and complex tones), listeners attempted to distinguish between different arrangements of repeated sequences consisting of the same three items, which were presented without any acoustic interactions or transitions involving contiguous sounds (Warren, 1974; Warren & Ackroff, 1976). These studies demonstrated that the different arrangements of nonverbal sounds could be discriminated with ease for item durations from 5 through

100 msec—yet the naming of orders was not possible within this range. It was suggested that permuted orders of brief items could be distinguished through the bonding of components to form temporal compounds possessing characteristic qualities, even though the component acoustic elements and their arrangements could not be identified. Thus, for an isomeric pair of temporal compounds consisting of identical components arranged in different orders, a listener might describe one arrangement of the nonverbal sounds as “bubbly” and the other as “shrill.” These qualitative differences served as the basis for accurate differentiation between different acoustic orders.

Vowel sequences and their verbal temporal compounds. Our subjects in Experiment 1 indicated that discrimination of permuted orders was accomplished in different ways at different item durations. When the item durations corresponded to single glottal pulses (10- and 12-msec vowels), listeners used nonverbal temporal compounds to distinguish between the permuted vowels. Thus, with these very brief durations, an individual might report, for example, that one order was characterized by a “dull” quality while the other order sounded “crisp.” However, perceptual organization into syllables and words (verbal temporal compounds) occurred at vowel durations from 30 through 100 msec. Within this durational range, a listener might say that one arrangement of vowels resembled or brought to mind repetitions of the word “kettle,” whereas the other arrangement sounded more like repetitions of “puddle”—this, despite the great phonetic differences between the actual stimuli and their lexical correlates. The specific word corresponding to a particular temporal arrangement varied from listener to listener.

It appeared desirable to study further the verbal organization of a succession of steady-state vowels into words, and Experiment 2 was undertaken in accordance with this purpose.

EXPERIMENT 2A: IDENTIFYING DIFFERENT ARRANGEMENTS OF 10-ITEM VOWEL SEQUENCES

Experiment 1 has shown that recycled sequences of steady-state vowels played loudly and clearly can be heard as coherent verbal utterances, and that different arrangements of the same vowels can be discriminated on the basis of their distinctive verbal organizations. Further informal observations indicated that roughly 30–80 msec/vowel was the optimal duration for hearing words. Experiment 2A was designed so that the characteristics of this vowel-word illusion could be examined using recycled sequences consisting of 10 40-msec vowels. The 400-msec duration of these sequences corresponded to that of words in normal conversation. During the experiment, listeners were presented with four recycled sequences, each having a different randomly determined vowel order, and they were instructed to use verbal organizations as a means

of identifying the different patterns on second presentation.

Method

Subjects. Thirty-two audiotometrically screened listeners (14 male and 18 female) were recruited from introductory psychology courses; they received either course credit or cash for participating. The screening procedure was the same as that described for Experiment 1.

Stimuli. For synthesis of the 10 vowel components, a Data Precision Co. Model 6100 Universal Waveform Analyzer, operating at a sampling rate of 40 kHz with 14-bit resolution, was used to excise single 5-msec glottal pulses from a male speaker's sustained productions (200-Hz voicing frequency) of 10 vowels (those in *heed*, *hid*, *head*, *had*, *hod*, *hawd*, *hood*, *hud*, *hoot*, and *herd*). The digitized glottal pulses were then iterated eight times, to produce 40-msec bursts that were judged by a panel of four trained listeners to be identifiable as the parent vowels. Linear ramps of 2.5 msec (0 dB minimum) were imposed upon the onset and offset of each vowel burst for suppression of transients, and the amplitude envelopes of the bursts were rescaled so that each would play back at the same level.

The 10 vowel bursts were sampled randomly without replacement and concatenated in digital form to create 48 10-item sequences (out of a total of factorial nine possible orderings). Digital-to-analog conversion and playback of the 400-msec sequences in recirculating form was accomplished using a Data Precision Co. Polynomial Waveform Synthesizer Model 2020-100 (40-kHz sampling frequency with 12-bit resolution). The analog playback of the recycling sequences was recorded on an Otari Model MTR 90-II 16-track recorder, with sequences to be presented on the same trial (4 sequences for each of the 12 trials) recorded in parallel on separate tracks. During the experiment, the output of the recorder was amplified by a Neotek Series I audio mixer and band-pass filtered from 50 to 8000 Hz with slopes of 115 dB per octave (Wavetek/Rockland Model 751A Brickwall Filter).

Procedure. The listeners were tested individually in an audiometric room, with the stimuli delivered at 70 dBA SPL through diotically wired TDH-49P headphones mounted in MX 41/AR cushions. The experimenter operated the Otari recorder (located outside the chamber) with a remote preset search-to-cue device. Switches on the audio mixer located inside the chamber permitted delivery of the output from the desired tracks of the recorder.

The listeners participated in 2 practice trials and 10 formal trials, with the 12 sets of sequences presented in the same order to all listeners. Each trial consisted of a learning phase and a test phase. During the learning phase, the listeners were presented successively with four sequences, and they were required to listen to the recycling vowel patterns until they could write down what the voice seemed to be saying. (For their transcriptions, the listeners used a response booklet with separate pages for each experimental trial.) It was explained that their written descriptions would provide a means of identifying the sequences during the test phase of the trial. Once the listener had provided written responses for each of the four sequences, the listener began the test phase, using a control box with buttons labeled A, B, C, and D. Each of the buttons could be used to deliver one of the four sequences presented during the learning phase of the trial, and the listener's task was to match the letter of each button with the previous verbal organization for that sequence. The listener did this by placing appropriately lettered cards beside the previous transcription.

The listeners were permitted to switch at will from one sequence to another during a trial's test phase, and they were given as much time as needed to complete the card-placing task. When matching was complete, the experimenter recorded the listener's response, provided feedback concerning accuracy, and began the next trial.

During the debriefing period that followed the 10th formal trial, the experimenter reviewed the transcriptions to verify pronunciation and asked general questions concerning the listener's responses.

Results

Despite the obvious initial doubt of most listeners that they could accomplish the experimental task, their perceptual organization of the recycling vowel sequences into syllables and words proved nearly effortless with little practice. The time required for initial verbal organization (that is, for writing down a description for a particular vowel sequence) decreased from an average of about 35 sec for the first practice sequence to an average of only 8 sec per sequence across the 10 formal trials. Furthermore, once formed during the learning phase of a trial, these perceptual organizations proved sufficiently distinct and stable to permit rapid and highly accurate identification of the different vowel orderings during the test phase. On the average, the listeners completed the four matches of the test phase in about 15 sec, and a majority of their responses were accurate even for the practice trials. Table 3 lists each trial separately and gives the numbers of listeners who identified correctly each of the four sequences for the individual trials.

The chance likelihood of correctly identifying all four sequences on a trial was 1/24, so each fully correct series of responses by a listener exceeded chance at the .05 level. As can be seen, listeners identified all four sequences with above-chance accuracy on most (better than 94%) of their attempts across the 10 formal trials, with no evidence of fatigue or interference due to earlier sequences.

For the 40 sequences presented in the formal trials, 35% of the listeners' responses were nonlexical syllables (which always followed the rules for phoneme clustering of English), and the remaining 65% were words and phrases. Interestingly, most listeners also reported that certain sequences were organized as two different words (e.g., "Frankie" and "go animal") that sounded as though they were produced simultaneously by voices differing in quality. Despite the fact that the sequences were presented in the same order to all listeners, there was very little inter-subject agreement in the forms reported for specific vowel orderings. Thus, although the verbal organizations were formed rapidly and were sufficiently stable to permit later recognition of sequences, they were also highly idiosyncratic—perhaps due in part to the fact that the se-

quences were played as endless loops with no initial and terminal components.

Experiments 1 and 2A have shown that verbal mediation in the discrimination of random vowel sequences can be very robust when differences in order are substantial. Experiment 2B was designed so that we might determine whether lexical matching could be extended to the discrimination of minimal differences in order.

EXPERIMENT 2B: DISCRIMINATION OF MINIMAL ORDER DIFFERENCES WITH TEN-ITEM VOWEL SEQUENCES

In the previous experiments, permuted orders of brief vowels produced distinct verbal organizations, but the differences in order were typically quite extensive: In the two contrasting three-item sequences used in Experiment 1 (ABCA... and ACBA...), each of the three pairwise orderings of vowels was reversed (AB vs. BA, BC vs. CB, and CA vs. AC), and in Experiment 2A, the 48 10-item sequences were drawn without constraint from a pool of 362,880 possible recycled orders. In Experiment 2B, listeners made ABX judgments (deciding whether the unknown X was the same as A or B) for 10-item vowel sequences, in which A and B differed only in the ordering of two contiguous vowels. The listeners also reported the basis for their discriminations for each trial.

Method

Subjects. Four subjects participated in the study. Subjects B.B. and J.B. were psychoacoustically trained listeners who had participated in preliminary observations with 10-item sequences. Listeners J.R. and K.R. were not psychoacoustically trained and had no prior experience with the stimuli employed in this study.

Stimuli. Each of the 48 sequences used in Experiment 2A was used as Sequence A of a contrasting pair. The B sequence of each pair was produced by interchanging the order of two randomly selected contiguous vowels of the 10-item A sequence. Analog playback of the B sequences was recorded on the same 16-track tape as had been used for Experiment 2A, with corresponding A and B sequences arranged in parallel. As in Experiments 1 and 2A, the stimuli were amplified, using an audio mixer, and band-pass filtered from 50 to 8000 Hz with slopes of 115 dB/octave.

Procedure. As in the earlier experiments, the listeners were tested individually in an audiometric room, with the stimuli delivered through headphones at 70 dBA SPL. They were provided with a three-button panel, which they used for switching between contrasting A and B stimuli and a third X stimulus that matched the sequence presented in either the A or the B channel. The listeners switched at will between the three signals (each recorded on a separate track) until they were satisfied that they had determined which signal matched X. After calling out either "A" or "B," they attempted to describe the basis for their discrimination. The listeners were aware that their ABX matches were being timed, and they received trial-by-trial feedback concerning their matching accuracy.

The listeners participated in a total of 16 sessions, with each session lasting about 20 min and involving judgments of 6 pairs of contrasting A and B sequences. Across the 16 sessions of each experiment, the 48 sequence pairs were presented twice to each listener, for a total of 96 judgments. Each listener received a different random ordering of stimuli for the first ABX judgments of the sequence

Table 3
Numbers of Listeners (out of 32) with Perfect Scores
(Correct Identification of Each of the Four
10-Item Vowel Sequences in a Trial) in Experiment 2A

	Practice Trials		Formal Trials									
	1	2	1	2	3	4	5	6	7	8	9	10
No. perfect scores*	22	27	28	31	30	31	28	31	32	32	27	32

* $p < .05$ for each perfect score.

Table 4
Accuracy and Response Times for ABX Judgments
of Recycled 10-Vowel Sequences in Experiment 2B
(A and B Sequences Differed in the Order
of a Single, Contiguous Pair of Vowels)

Listener	No. Correct (out of 96)*	Response Times (in sec)		
		Median	Q ₁	Q ₃
B.B.	92	34.5	25.0	51.0
J.B.	94	50.5	30.0	107.5
J.R.	94	72.0	41.0	114.0
K.R.	94	42.0	28.0	68.5

*Accuracy scores for all listeners exceeded chance ($Z \geq 8.98$, $p < .0001$).

pairs, and this order was repeated for the listener upon second presentation of the stimuli, so that the two judgments for each contrast were separated by judgments of the remaining 47 sequence pairs.

Results

The number of correct responses (out of 96) and the median response times for judgments of each listener are presented in Table 4. As is shown, overall matching accuracy was well above chance for all listeners, with the percentage of correct responses ranging from about 96% to 98%.

The listeners' trial-by-trial reports concerning the nature of their discriminations indicated that, although they attributed some discriminations to contrasting nonverbal characteristics (typically, differences in rhythmic complexity), most of their judgments were based on differences in verbal organizations. These occasionally corresponded to pseudowords, but more often to real words (e.g., "valuable" vs. "technical"). Most interestingly, although there was little agreement across listeners in the verbal forms evoked by specific vowel sequences, there was substantial consistency within listeners: In 52% of the cases in which listeners reported specific words upon first presentation of a contrasting pair of sequences, they reported the same word or words on second presentation of the sequences. This repetition of responses occurred in spite of the fact that successive judgments of the same stimuli were separated by several days and by interpolated judgments of the remaining 47 sequence pairs. Thus, although the verbal correlates of these monotone vowel patterns were again found to be highly idiosyncratic, they were also remarkably stable.

Discussion

In studies with 10-item sequences of nonverbal sounds, it has also been found that minimal changes in 10-item sequences can be discriminated. Watson and his co-workers employed sequences of 10 or more brief sinusoidal tones in experiments on the ability to make fine discriminations (e.g., the ability to detect a change in the frequency of a single tone) within complex "word-length" patterns (see Watson, 1987, for a review). In these studies, in which contrasting sequences were presented as single statements, it was found that listeners usually required

many hours of training before they could accomplish discrimination. However, Bashford and Warren (1988) have reported that when sequences of 10 tones are recycled, the discrimination of fine changes is very much easier and can be accomplished in less than 1 min in an ABX discrimination task. They found performance with minimal changes (inverting the order of 2 of the 10 tones) to be only slightly poorer than that observed with recycled 10-vowel sequences. Hence, although perception in a "speech mode" (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) is employed for sequences of vowels, fine discrimination can be accomplished with non-linguistic sequences as well. Bashford and Warren (1988) also reported that sequences need not involve successions of discrete sounds for successful discrimination. Noise "sequences" were constructed by sampling from a catalogue of 10 40-msec segments that had previously been excised from white noise. When the segments were abutted to form a loop, the recycled "sequence" resembled a repeated 400-msec segment of noise that lacked the succession of discrete sounds characteristic of sequences of tones and vowels. Interchanging the order of two contiguous 40-msec noise segments resulted in a discriminable change—although ABX judgments did take about twice as long as those with the recycled sequences consisting of 10 40-msec tones or vowels.

Let us look more closely at the linguistic organization occurring within sequences of vowels in Experiments 1, 2A, and 2B. How is it that syllables and words are heard with a succession of steady-state vowels, despite the great differences between the phonetic compositions of the stimuli and the forms reported?

We hypothesize that the organization of our sequences of loud and clear vowels into syllables and words reflects shifts in perceptual criteria produced by repetition. The *criterion shift rule*, which has been proposed for judgmental processing in general, considers that the criteria used for evaluating stimuli and events are displaced in the direction of simultaneous or recently experienced values (Warren, 1985). When applied to psycholinguistics, this effect can produce changes in the perceptual boundaries of phonemes following exposure to repeated syllables. While there is considerable controversy concerning the processes responsible for these boundary shifts (for discussion, see Diehl, Kluender, & Parker, 1985), there is agreement that the changes that do occur move the acoustic boundaries delineating particular phonemes toward a closer correspondence with the iterated stimulus. This shifting of criteria may be considerably greater when repetition is continuing (as in the present experiments) than after repetition ceases (as is typically the case in studies measuring the extent of category boundary shifts). In the present experiments, it appears that the continuing repetition of a loud and clear sequence of steady-state vowels changed the acoustic requirements for recognition of a syllable or word to the point at which the stimulus itself could be perceived as a particular utterance by a speaker.

The perceptual matching of a repeated vowel sequence to a particular verbal form may be facilitated not only by a criterion shift, but also by the splitting of the stimulus into two simultaneous percepts. Recall that typically an iterated vowel sequence splits into two concurrent forms—usually, two voices with different pitches or qualities, which repeat different things at the same time (although sometimes a single verbal form is heard, accompanied by a nonverbal sound). We suggest that matching of the auditory input to the particular patterns (or templates) required for perception of syllables or words involves separation of the signal into two fractions. One fraction is matched to the template corresponding to a syllable or word (as modified by a repetition-induced criterion shift). The other fraction corresponds to the residue remaining after subtraction of the components of the auditory input that are used for this match. This residue can appear as a nonlinguistic noise, or it may be matched to a second linguistic template and thus heard as a different voice repeating some other utterance. The process of synthesizing an auditory signal through subtraction of the appropriate components from a louder sound has been called *auditory induction* (Warren, 1984; Warren, Obusek, & Ackroff, 1972). In conjunction with repetition-induced shifts in acoustic criteria defining linguistic templates, auditory induction could facilitate the matching of vowel sequences to syllables and words, by permitting the segregation of spectral components corresponding to these modified templates.¹

It would be of interest to determine the correspondence of individual speech sounds forming the illusory words to the vowels actually present at that time. Preliminary experiments have shown that the mapping of perceptual phonemes to acoustic phonemes can be accomplished, but not through methods that might appear to be the most obvious. Placing an acoustic marker such as a click in one of the vowels does not work, since clicks (and other extraneous sounds) are mislocalized in speech (Ladefoged, 1959; Warren & Obusek, 1971). Increasing the intensity of a vowel appreciably and then listening for a corresponding increase in the level of speech sounds in the illusory word does not work, because the illusory word usually continues to be heard, and the increased intensity results in one's hearing the vowel veridically—but as an extraneous sound that cannot be localized in the word. Deleting a vowel and listening for the disappearance of a portion of the illusory word does not work, because the illusory word can change to another form. However, a method for phoneme mapping of recorded speech employed in earlier studies (Warren, 1971; Warren & Sherman, 1974) does appear to work quite well. When the repeated sequence of vowels is abruptly terminated, the illusory word (or words) also stops suddenly, and it is easy to perceive the last speech sound heard. By systematically changing the point of termination of the sequence of vowels, one can map the perceptual phonemes to the acoustic phonemes. Further work employing this procedure is in progress.

SUMMARY AND CONCLUSIONS

Experiment 1 shows that repeated sequences consisting of different arrangements of the same three vowels can be distinguished either through naming the order of components (for item durations greater than 100 msec) or by means of recognition of patterns through temporal compound identification (for durations from 10 through 100 msec). Perception in a speech mode occurred for items from 30 through 100 msec, allowing permuted orders to be discriminated through perception of different verbal organizations for the different arrangements. Nonverbal temporal compounds permitted the discrimination of different arrangements for vowels briefer than 30 msec. In Experiments 2A and 2B, we examined the speech mode of perception further, by employing complex repeated sequences of 10 40-msec vowels. The recognition of different arrangements was accomplished readily through verbal mediation, even for the minimal changes in order produced by interchanging the position of two contiguous items. The vowel sequences were heard as a single utterance plus a noise, or as two concurrent utterances produced by distinctly different voices.

It was hypothesized that two mechanisms are involved in the illusory perception of words with repeated sequences. The syllabic or lexical templates employed for verbal recognition were temporarily warped into a closer resemblance to the repeated stimulus through repetition-induced criterion shifts, and matching of the stimulus to the template was then completed by extracting components needed for the match from the auditory input. This perceptual splitting of the stimulus (which also occurs during phonemic restoration) produced a residue that was either perceived as an extraneous sound accompanying the illusory verbal organization or organized into a second verbal form heard along with the first.

It is of interest that studies with animals other than humans have shown that, although the animals can discriminate between different arrangements of brief sounds, they fail when the task requires the remembering of sounds for more than a few seconds. As is discussed below, this difference between the performance of humans and other animals has suggested how speech perception might have evolved from auditory skills possessed by our prelinguistic ancestors.

On the basis of a literature survey of studies demonstrating that cats, chinchillas, and monkeys can be taught to recognize not only isolated phonemes, but also monosyllables, the suggestion has been made that the mechanisms employed by humans for speech perception have evolved through the elaboration of an ability to recognize overall patterns (or temporal compounds) that we share with other animals (Warren, 1982, 1988). In addition to the animal studies involving sequences of speech sounds, other experiments involving periodic sounds and noises have shown that dolphins (Thompson, 1976) and monkeys (Dewson & Cowey, 1969) can be taught to discriminate between pairs of brief sounds ar-

ranged in different orders. However, successful discrimination could be accomplished only when the sequences were brief; when the task required that these animals remember the identity of the first sound for 2 sec or more before hearing the second sound, the task became impossible (for further discussion, see Warren, 1982, pp. 137-138). It seems that discrimination between sequences with long separation between items requires a mechanism that is lacking in other animals but available to humans. This mechanism appears to involve verbal encoding, so that, for items separated by more than a few seconds, linguistic characterizations (rather than the memory of the sounds themselves) are stored to serve as the basis of discrimination.

For recognition of sequences with brief item durations (such as speech), neither humans nor other animals need identify the order of components or even the components themselves. Only temporal compounds need be recognized. Although listeners may be able to name the ordered series of phonemes corresponding to a word, this analytical description does not necessarily imply that the components themselves are perceived. Thus, Brubaker and Warren (1988) have demonstrated that listeners can readily learn to name the order of acoustic phonemes corresponding to words that are perceived, even when these words have phonetic transcriptions that do not correspond to the acoustic-phonetic components. They used recycled sequences of three vowels (as in Experiment 1). Their subjects were first presented with the two possible arrangements of the vowels at item durations of a few hundred milliseconds (permitting easy identification of order). They then heard these sequences at item durations that were decreased in a regular fashion down to values well below the threshold of 100 msec reported for identification of order with recycled sequences of vowels (Dorman et al., 1975; Thomas et al., 1971). At no time were the subjects ever told the actual phonemes or their orders. Through a series of successive generalizations, the subjects continued to identify accurately the constituent vowels in their proper orders, even though, as in the present study, the words heard at brief item durations did not have phonetic transcriptions corresponding to the acoustic phonemes actually present in the stimulus. It was concluded that the perception of syllables and words did not involve a "bottom up" or prior identification of an ordered arrangement of phonetic components. Rather, the identification of the acoustic phonemes and their orders required the mediation of a prior verbal organization.²

The recognition of lexical items in connected discourse, of course, consists of more than just the factors described above. Syntactic, semantic, and pragmatic rules come into play with lexical aggregates, and these emergent higher level processes can in turn influence word recognition. However, experiments involving perception of isolated words and phrases (as in the present study) can provide information concerning some of the flexible and opportunistic mechanisms used for the early stages of verbal processing.

REFERENCES

- BASHFORD, J. A., JR., & WARREN, R. M. (1988). Discrimination of recycled word-length sequences. *Journal of the Acoustical Society of America*, **84**(Suppl. 1), S154.
- BRUBAKER, B. S., & WARREN, R. M. (1988). Learning to identify phonemic orders. *Journal of the Acoustical Society of America*, **84**(Suppl. 1), S154.
- COLE, R. A., & SCOTT, B. (1973). Perception of temporal order in speech: The role of vowel transitions. *Canadian Journal of Psychology*, **27**, 441-449.
- CULLINAN, W. L., ERDOS, E., SCHAEFER, R., & TEKIELI, M. E. (1977). Perception of temporal order of vowels and consonant-vowel syllables. *Journal of Speech & Hearing Research*, **20**, 742-751.
- DEWSON, J. H., III, & COWEY, A. (1969). Discrimination of auditory sequences by monkeys. *Nature*, **222**, 695-697.
- DIEHL, R. L., KLUENDER, K. R., & PARKER, E. M. (1985). Are selective adaptation and contrast effects really distinct? *Journal of Experimental Psychology: Human Perception & Performance*, **11**, 209-220.
- DIVENYI, P. L., & HIRSH, I. J. (1978). Some figural properties of auditory patterns. *Journal of the Acoustical Society of America*, **64**, 1369-1385.
- DORMAN, M. F., CUTTING, J. E., & RAPHAEL, L. J. (1975). Perception of temporal order in vowel sequences with and without formant transitions. *Journal of Experimental Psychology: Human Perception & Performance*, **104**, 121-129.
- LADEFOGED, P. (1959). The perception of speech. In *National Physical Laboratory Symposium No. 10: Mechanisation of thought processes* (pp. 309-417). London: Her Majesty's Stationery Office.
- LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.
- NICKERSON, R. S., & FREEMAN, B. (1974). Discrimination of the order of the components of repeating tone sequences: Effects of frequency separation and extensive practice. *Perception & Psychophysics*, **16**, 471-477.
- REPP, B. H. (1989). Phone restoration. *Journal of the Acoustical Society of America*, **85**(Suppl. 1), S137. (Abstract No. DDD9)
- SKINNER, B. F. (1936). The verbal summator and a method for the study of latent speech. *Journal of Psychology*, **2**, 71-107.
- TERANISHI, R. (1977). Critical rate for identification and information capacity in hearing system. *Journal of the Acoustical Society of Japan*, **33**, 136-143.
- THOMAS, I. B., CETTI, R. P., & CHASE, P. W. (1971). Effect of silent intervals on the perception of temporal order for vowels. *Journal of the Acoustical Society of America*, **49**, 84.
- THOMAS, I. B., & FITZGIBBONS, P. J. (1971). Temporal order and perceptual classes. *Journal of the Acoustical Society of America*, **50**, 86-87.
- THOMAS, I. B., HILL, P. B., CARROLL, F. S., & GARCIA, B. (1970). Temporal order in the perception of vowels. *Journal of the Acoustical Society of America*, **48**, 1010-1013.
- THOMPSON, R. K. R. (1976). *Performance of the bottlenose dolphin (Tursiops truncatus) on delayed auditory sequences and delayed auditory successive discriminations*. Unpublished doctoral dissertation, University of Hawaii.
- WARREN, R. M. (1968). Relation of verbal transformations to other perceptual phenomena. *Conference Publication No. 42: Institution of Electrical Engineers* (Suppl. 1), 1-8.
- WARREN, R. M. (1971). Identification time for phonemic components of graded complexity and for spelling of speech. *Perception & Psychophysics*, **9** (4), 345-349.
- WARREN, R. M. (1972). Perception of temporal order: Special rules for initial and terminal sounds of sequences. *Journal of the Acoustical Society of America*, **52**, 67.
- WARREN, R. M. (1974). Auditory temporal discrimination by trained listeners. *Cognitive Psychology*, **6**, 237-256.
- WARREN, R. M. (1982). *Auditory perception: A new synthesis*. New York: Pergamon.

- WARREN, R. M. (1983). Multiple meanings of "phoneme" (articulatory, acoustic, perceptual, graphemic) and their confusions. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 9, pp. 285-311). New York: Academic Press.
- WARREN, R. M. (1984). Perceptual restoration of obliterated sounds. *Psychological Bulletin*, **96**, 371-383.
- WARREN, R. M. (1985). Criterion shift rule and perceptual homeostasis. *Psychological Review*, **92**, 574-584.
- WARREN, R. M. (1988). Perceptual bases for the evolution of speech. In M. E. Landsberg (Ed.), *The genesis of language* (pp. 101-110). Berlin: Mouton de Gruyter.
- WARREN, R. M., & ACKROFF, J. M. (1976). Two types of auditory sequence perception. *Perception & Psychophysics*, **20**, 387-394.
- WARREN, R. M., & BYRNES, D. L. (1975). Temporal discrimination of recycled tonal sequences: Pattern matching and naming of order by untrained listeners. *Perception & Psychophysics*, **18**, 273-280.
- WARREN, R. M., & OBUSEK, C. J. (1971). Speech perception and phonemic restorations. *Perception & Psychophysics*, **9** (3B), 358-362.
- WARREN, R. M., & OBUSEK, C. J. (1972). Identification of temporal order within auditory sequences. *Perception & Psychophysics*, **12**, 86-90.
- WARREN, R. M., OBUSEK, C. J., & ACKROFF, J. M. (1972). Auditory induction: Perceptual synthesis of absent sounds. *Science*, **176**, 1149-1151.
- WARREN, R. M., OBUSEK, C. J., FARMER, R. M., & WARREN, R. P. (1969). Auditory sequence: Confusion of patterns other than speech or music. *Science*, **164**, 586-587.
- WARREN, R. M., & SHERMAN, G. L. (1974). Phonemic restorations based on subsequent context. *Perception & Psychophysics*, **16**, 150-156.
- WARREN, R. M., & WARREN, R. P. (1970). Auditory illusions and confusions. *Scientific American*, **223** (December), 30-36.
- WATSON, C. S. (1987). Uncertainty, informational masking, and the capacity of immediate memory. In W. A. Yost & C. S. Watson (Eds.), *Auditory processing of complex sounds* (pp. 267-277). Hillsdale, NJ: Erlbaum.
- WICKELGREN, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, **76**, 1-15.

NOTES

1. Another example of linguistic auditory induction is given by the *phonemic restoration effect*, in which contextually appropriate fragments of speech are synthesized from the substrate furnished by a louder sound of appropriate spectral characteristics (for a detailed discussion, see Warren, 1984). In keeping with induction theory, Repp (1989) has reported that spectral components corresponding to the restored phoneme are subtracted from an interpolated noise.
2. It has been suggested by Warren (1983) that there are four rather different uses of the term *phoneme* (acoustic, articulatory, graphemic, and perceptual), and that confusion has resulted from employing the same term for different entities. Warren argued that the existence of phonemes as units entering into the perceptual processing of discourse lacks direct experimental support, and that the treatment of perceptual phonemes in the literature is often confounded with acoustically based phonemes and with articulation-based phonemes.

(Manuscript received September 18, 1989;
revision accepted for publication November 20, 1989.)