Journal of Experimental Psychology:
Human Perception and Performance
2002, Vol. 28, No. 1, 150–162

0096-1523/02/$5.00 DOI: 10.1037//0096-1523.28.1.150

Linking Verbal Transformations to Their Causes

Mark A. Pitt and Lisa Shoaf
Ohio State University

The verbal transformation effect (VTE) is a perceptual phenomenon in which listeners report hearing illusory utterances when a spoken word is rapidly repeated for an extended period of time. The cause of the illusion was investigated by identifying regularities across the transformations that listeners reported and then testing hypotheses about the cause of those regularities. Variants of the standard transformation paradigm were used across 3 experiments to demonstrate that perceptual regrouping of the elements in the repeating utterance is 1 cause of the VTE. Findings also suggest that regrouping is influenced by whether the stimulus is perceived as speech or as nonspeech.

Within the information-processing approach to perception, one method of inquiry is to investigate the causes of illusions, which are viewed as windows into the processes that operate during veridical perception. That is, illusions are temporary lesions (Warren, 1968) of the perceptual system that reveal its inner workings.

One illusion whose cause continues to elude researchers despite years of study is the verbal transformation effect (VTE; Warren, 1961b). The VTE is experienced while listening to a word repeat at a fast rate (e.g., once per half second). Initially, the veridical percept (i.e., the intended word) is heard, but at some point illusory utterances begin to be perceived. For example, a subset of transformations reported in response to the word *ripe* includes “right,” “ride,” “rife,” “life,” “bright,” “rape,” and “wife” (Warren, 1961b). Verbal transformations are so vivid that participants believe that the stimulus changes along with the percept, making the VTE a very compelling illusion. In general, a few specific transformations are heard by many listeners, and each listener also reports idiosyncratic transformations. Although some transformations can be small phonetic deviations from the veridical percept, as in the examples above, others are positively weird (e.g., reporting “florist” when listening to *trice*).

Investigations of the illusion have largely focused on the dynamics of switching among transformations. In recent work, Ditzinger, Tuller, and Kelso (1997) found evidence that switching is not always random but can involve rapid and long alternations between pairs of transformations, suggesting that there may be a coupling mechanism that underlies the illusion much like that responsible for alternating between views of a reversible figure.

Warren (1968, 1976; Warren & Meyers, 1987) considers the VTE, like phonemic restoration, to reflect the operation of processes devoted to the perceptual organization and interpretation of speech. Switching among transformations is caused by two processes whose function is to repair or reinterpret the speech signal when it does not make sense: satiation and a criterion shift. Repeated presentation of an utterance causes its memory representation to satiate. Simultaneously, the criteria used to categorize the stimulus shift so that another representation is likely to be deemed a better match to the input than the original representation (Warren, 1985). When this occurs, a transformation is heard. The processes of satiation and criterion shifting then begin anew, repeating themselves throughout the presentation of the stimulus. MacKay, Wulf, Yin, and Abrams (1993) and Natsoulas (1965) have also invoked the concept of satiation to account for the illusion. Lexical and sublexical (e.g., syllabic) levels of representation have been suggested as the loci of such effects.

In contrast to investigations that have examined switching between transformations, comparatively few studies have explored why specific transformations are reported. The contents of transformations (e.g., phonetic, lexical, semantic) reported by listeners may be an equally valuable source of information about the cause of the illusion. For example, Goldstein and Lackner (1973) and Clegg (1971) compared the phonetic content of the transformations that listeners reported with that of the veridical percept. There was a tendency for the veridical phoneme to be replaced by a phonetically similar one (e.g., /t/ for /k/; /ɛ/ for /æ/), suggesting that the underlying perceptual process induces a slight shift in categorization, which is what might be expected if satiation were involved. Similarly, Chalikia and Warren (1991, 1994; Warren, Healy, & Chalikia, 1996) have successfully used this approach to study phonemic transformations, which are illusory verbal percepts that are heard when listening to repeating strings of concatenated vowels.

Detailed analyses of what listeners report have the potential to reveal regularities in perception that can provide clues about the processes causing transformations. Such analyses can link specific types of transformations to specific perceptual processes. Making such a connection is vital, as the problem of not understanding why specific transformations are generated is one reason the illusion is so mysterious and difficult to penetrate. If researchers can discover these connections, the illusion should become more tractable, potentially making transformations predictable (if not in terms of

Mark A. Pitt and Lisa Shoaf, Department of Psychology, Ohio State University.

This work was supported by National Institute on Deafness and Other Communication Disorders, National Institutes of Health Grant R29 DC01774. We thank Lyn Canterbury and Kalyani Subramaniam for help in scoring data, and we thank three reviewers for helpful feedback on an earlier version of the article. We also thank the generous support of The Center for Research in Computing and the Arts at the University of California, San Diego, which provided equipment and facilities for this project during the fall of 1998.

Correspondence concerning this article should be addressed to Mark A. Pitt, Department of Psychology, Ohio State University, 1885 Neil Avenue, Columbus, Ohio 43210. E-mail: Pitt.2@osu.edu

the exact utterances, then certainly in terms of properties those utterances are likely to possess). We adopted this approach to studying the VTE in the present study.

The starting point for the present investigation grew out of an analysis of transformations reported in a study examining lexical influences in the VTE (Shoaf & Pitt, in press). In some of the transformations reported to the words *skunk* and *skin*, /s/ was missing, and the transformation began with /g/ (e.g., /gʌŋk/, /gɪn/, /gɛn/). We hypothesized that these transformations were caused by the perceptual regrouping of the acoustic elements that made up the word. Specifically, the repetitive presentation of the words caused the high-frequency frication corresponding to /s/ to split off from the remainder of the word and form a separate perceptual stream containing only /s/. Listeners heard the remainder as the foreground percept and reported it as the transformation. The reason /g/ was reported instead of /k/ was that regrouping eliminated the influence that /s/ has on perception of the following stop consonant (/k/). When spoken syllable initially (e.g., /ka/), /g/ and /k/ are distinguished in part by the presence of aspiration in /k/ but not /g/ (Delattre, Liberman, & Cooper, 1955). Aspiration is absent, however, when /k/ is produced after /s/, as in *skin*. When /s/ is removed from such a word, listeners will tend to report /g/ instead of /k/ because the stop consonant is now acoustically similar to /g/. Thus, the fact that listeners reported /g/-initial transformations to words like *skin* suggests /s/ split off from the remainder of the utterance.

Support for a perceptual regrouping explanation of these transformations has come from multiple sources that have demonstrated that speech can split into multiple streams. Most relevant to the preceding observation is that streaming has been identified as the culprit in producing unwanted methodological artifacts in other paradigms in which a recycling utterance is used (Cooper, Whalen, & Fowler, 1986; de Jong, 1994). Perhaps most relevant to the current study is an exchange between researchers over the likelihood that streaming contributed to the outcome of a selective adaptation experiment in which /spa/ (the adaptor) was repeated continuously for an extended period (Diehl, Kluender, & Parker, 1985; Diehl, Parker, & Kluender, 1985; Sawusch & Jusczyk, 1981; Sawusch & Mullennix, 1985). Diehl et al. argued that /s/ split off from the remainder of the syllable, causing listeners to hear /b/, not the intended phoneme, /p/.

Studies have also identified conditions that cause speech to stream. Using a repetitive presentation paradigm, Cole and Scott (1973; see also Cullinan, Erdos, Schaefer, & Tekieli, 1977; Dorman, Cutting, & Raphael, 1975; Lackner & Goldstein, 1974) found that adjacent phonemes cohere and form a single perceptual stream much better when there are formant transitions between segments than when there are not. Without transitions, the phonemes cohered poorly with one another, causing listeners to lose track of their temporal order, a hallmark characteristic that multiple streams were perceived (Bregman & Campbell, 1971; Warren, Obusek, Farmer, & Warren, 1969). Chalikia and Warren (1994) showed that repeating vowel sequences break into parallel streams containing nonoverlapping frequency regions of the input, suggesting that segments similar in frequency group together. Finally, when presented in close temporal succession, recycling words can be perceptually resegmented whereby a boundary phoneme is heard word-finally instead of word-initially (and vice versa), resulting in another word (e.g., *fly* → *life*, *leap* → *plea*; MacKay et al., 1993; Reisberg, Smith, Baxter, & Sonenshine, 1989). In effect, the phoneme is regrouped with the other end of the word. Differences in frequency and temporal proximity between the vowel and the final consonant, and between the final and initial consonants, probably contribute to resegmentation.

The strength of the preceding evidence warranted a formal test of the proposal that auditory regrouping is one cause of verbal transformations. In Experiments 1–3 we investigated the plausibility of this proposal.

Experiment 1

If perceptual regrouping causes verbal transformations, then listeners' percepts should exhibit characteristics that are indicative of streaming, and the frequency with which such streaming-based transformations are reported should depend on the acoustic properties of the stimuli. One characteristic of perceptual regrouping is that the percept splits into a foreground stream and one or more background streams (Brochard, Drake, Botte, & McAdams, 1999). Transformations suspected of being due to streaming should therefore be accompanied by reports of multiple streams. In debriefing sessions in Shoaf and Pitt (in press), some listeners described what they heard in these terms, providing preliminary evidence that streaming does cause verbal transformations. This prompted us to also wonder whether listeners could selectively attend to each stream and identify its contents.

If the elements of the recycling utterance do split into multiple streams, then, on the basis of what is known about the conditions that promote speech to stream, we should be able to exert some control over what portion of the stimulus will split off and form its own stream. Specifically, adjacent phonemes that are similar to each other in ways that promote coherence should stream less easily and less often than those that do not. The above literature on the conditions that cause speech to stream suggests that those phonemes should cohere that occupy similar frequency regions, possess clear formant transitions, and are temporally contiguous (i.e., there are no gaps between the acoustic elements that constitute the phonemes). For example, continuants such as /w/ might be expected to cohere reasonably well with a following vowel because their energy regions are similar, formant transitions connect the two, and they are temporally contiguous. Fricatives and affricates should cohere less well because the acoustic energy of the fricative will be much higher than the vowel formants, and the formant transitions are less salient. Initial voiceless stops also might not cohere well with the vowel. Aspiration of the stop is higher in frequency and weaker in amplitude than the vowel formants, and the stop burst is temporally separated from the onset of the vowel (i.e., voicing).

In Experiment 1, we tested these ideas. In addition to reporting verbal transformations, listeners were instructed to report the number of streams heard and their contents. The stimuli were CVC pseudowords (C = consonant, V = vowel) created so that the consonants varied in their likelihood of splitting off and forming a separate stream. Consonants varying in cohesiveness with the vowel were paired together in various combinations to yield three stimulus conditions: *intact*, in which neither consonant was expected to stream frequently; *final*, in which the last consonant was expected to stream but not the first; and *initial plus final* (I + F), in which both the initial and final consonants could split off from the vowel. If the perceptual regrouping hypothesis is correct, transformations caused by streaming should be identifiable by listeners reporting two streams, each containing a portion of the veridical percept. Moreover, the frequency with which a CVC yields streaming transformations as well as which segment splits off should vary as a function of the acoustic characteristics of the utterance.

Method

Participants. Thirteen undergraduates from an introductory psychology course participated for course credit. All were native English speakers and reported normal hearing.

Apparatus. Participants were seated in individual sound-attenuated rooms. A microcomputer presented stimuli over headphones at a comfortable listening level. Verbal responses were recorded onto one track of a cassette tape from a microphone that was situated in front of the participant. A button box connected to the microcomputer collected manual responses.

Stimuli. Six CVC pseudowords served as stimuli. Pseudowords were used instead of real English words to minimize lexical influences, which have been found to affect listeners' reports (Natsoulas, 1965; Shoaf & Pitt, in press; Warren, 1961b). CVCs were used instead of more phonologically complex utterances to make the cause of the transformations as tractable as possible.

There were two stimuli in each of the three conditions (intact: /lom/, /wɛm/; final: /lodʒ/, /wɛtʃ/; I + F: /podʒ/, /pɛtʃ/). The affricates /dʒ/ and /tʃ/ and the voiceless stop /p/ were used as the consonants that would be most likely to stream from the utterance because their acoustic properties differ greatly from the adjacent vowel in periodicity, frequency, and, in the case of /p/, temporal proximity. The consonants /l/, /m/, and /w/ were used as those that would be less prone to stream because they are periodic, similar to the vowel in frequency, and provide smooth formant transitions into and out of the vowel. The same two vowels (/o/, /ɛ/) were used across the three stimulus conditions to facilitate comparisons across conditions. An *initial* condition, which would have consisted of an utterance-initial affricate or stop and an utterance-final liquid or nasal (e.g., /pɛm/), was not included because the optimum number of trials in the experiment would have been exceeded. In past work, we have found that too many participants became fatigued trying to maintain a high level of alertness and concentration after six or seven trials (2–3 min each) and thus stopped reporting transformations.

Stimuli were recorded onto digital audiotape by Mark A. Pitt and then digitally transferred to the hard disk of a PC (downsampled from 48 kHz to 16 kHz, 16-bit quantization, 7.6-kHz cutoff), where they were edited and saved as individual files. Mean stimulus duration in the intact and I + F conditions was similar (intact = 470 ms; I + F = 495 ms) and slightly shorter than that in the final condition (543 ms).

Procedure. The experiment began with a lengthy introduction during which listeners were familiarized with the concepts of perceptual regrouping and the VTE. Listeners were first presented with visual and auditory displays demonstrating perceptual regrouping of elements. The auditory examples consisted of sequences of tones that alternated between two frequencies. Listeners pressed a button labeled *one group* when the tones fused to form a single stream and another button labeled *two groups* when tones split into separate streams. Participants did not proceed to the next part of the experiment until they could reliably perform this task.

Next, listeners participated in a five-trial practice session in which they were introduced to the VTE first and then taught to report transformations. They were told to listen carefully to the repeating utterance, and if the stimulus changed into another utterance, even if it changed into one they had heard previously, they were to report the change into the microphone. Prior to the last two trials, the instructions were augmented to provide information about perceptual regrouping. This was accomplished by integrating into each trial the button-press response made to the tone sequences. On hearing a transformation, listeners were instructed to report the transformation, to indicate whether it formed one or two streams by pressing the button labeled *one group* or *two groups*, and then to report the contents of the other stream. If the contents of the other stream did not sound like speech, listeners were to describe it as best as possible.

In the test session, the six stimuli were presented for 250 repetitions each (0-ms interstimulus interval [ISI]) in one of two counterbalanced orders, making the design completely within subjects. The experiment lasted 1 hr.

Results

Listeners had little difficulty performing the two tasks. On only 1% of the trials did button-press responses mismatch the number of verbal reports made (e.g., pressing *two groups* and reporting the contents of only one stream). These trials were not included in the analysis. This low error rate is not simply a result of there being few opportunities for errors, as listeners reported hearing the stimuli form two streams slightly more often than they reported hearing one (60% vs. 40%, respectively). These data demonstrate that the experiment was successful in inducing the perception of multiple streams. The analyses discussed below examined how regrouping differed across the three stimulus conditions (intact, final, and I + F) and the phonetic contents of these streams.

Transformations due to regrouping (dubbed *streaming transformations*) were defined as instances in which listeners pressed the *two groups* button and reported two percepts, one being the transformation and the other being the contents of the second stream. As mentioned earlier, 60% of the responses exhibited this characteristic. The remainder were reports of the veridical percept (28%) or transformations that were not due to regrouping of the elements in the utterance (12%). These latter transformations consisted of consonant and vowel substitutions and insertions. Examples include /bioɛm/ and /gwɛm/ for /wɛm/, /blom/ and /loam/ for /lom/, /duɛt/ and /skuɛt/ for /wɛtʃ/, /pold/ for /podʒ/, and /loʌn/ for /lodʒ/. For additional examples of the range of transformations listeners report in various phonetic contexts, see Goldstein and Lackner (1973), Warren (1961a, 1961b), Warren and Meyers (1987), and Clegg (1971).

The proportion of streaming transformations in the three stimulus conditions is shown in the first column of Table 1. Streaming responses occurred almost half of the time in the intact condition, more frequent than we anticipated given that the consonants were expected to cohere with the vowel. Nevertheless, streaming transformations were reliably less frequent in the intact than in the final and I + F conditions, $\chi^2(2, N = 295) = 10.05$, $p < .01$, which did not differ reliably from each other. These data demonstrate that stimulus properties that promote perceptual regrouping lead to more streaming transformations.

The values in columns 3 and 4 of Table 1 show the preceding data subdivided as a function of whether the initial (C_1) or final (C_2) consonant split off and was reported as belonging to the less salient (background) stream. Because the consonants of the intact-condition stimuli were expected to adhere to the vowel, there were no expectations as to which consonant would split off most often. There was a bias for C_2, /m/, to form its own stream far more often than C_1, leading to transformations such as /wɛ/ and /wæ/ for /wɛm/, with /m/ forming the background stream. We found that /lom/ behaved similarly, with /lo/ and /la/ being the most frequent streaming transformations and /m/ reported in the background.

Of more interest are the results in the final and I + F conditions, where differences were anticipated. In the final condition, only C_2 split off, which demonstrates that VC cohesion was much weaker

Table 1
Transformation Measures in the Three Conditions of Experiment 1

	Proportion of streaming transformations			Mean no. of repetitions		
Condition	Overall	Initial location	Final location	To first streaming transformation	When one stream reported	When two streams reported
Intact (/lom/, /wɛm/)	.47	.17	.83	64	59	75
Final (/lodʒ/, /wɛtʃ/)	.67	.00	1.00	28	30	108
I + F (/podʒ/,/pɛtʃ/)	.63	.21	.79	34	34	70

Note. I + F = initial plus final.

than CV cohesion. For /wɛtʃ/, the most frequent streaming transformations were /wɛ/ and /wɛt/, with /tʃ/ being reported in the background stream for both of these. In addition, /lodʒ/ yielded similar CV transformations, with /lo/ and /lod/ being the typical transformations and /tʃ/ or /dʒ/ being typical background percepts.

As in the intact and final conditions, there was a bias in the I + F condition for C_2 to stream off most frequently. However, as predicted, C_1 did show evidence of streaming, being reported in the foreground stream one fifth of the time. When C_2 streamed, reports resembled those in the final condition. For /pɛtʃ/, /pɛt/ and /pɛ/ were reported as transformations, and /tʃ/ formed the background stream. For /podʒ/, /pod/ and /po/ were the most frequent transformations, and /tʃ/ and /dʒ/ were reported in the background stream. When /p/ streamed, transformations exhibited more variability. In a few instances, /ɛtʃ/ and /odʒ/ were reported as transformations, with /p/ completely disappearing into the background. However, the majority of transformations were /h/-initial variants of these, such as /hɛtʃ/, /hɛt/, and /hodʒ/, as if a portion of the /p/ aspiration adhered to the VC and was heard as /h/.

There were also a few transformations in the I + F condition that suggest both consonants streamed from the vowel simultaneously. For example, /hɛt/ was reported as a transformation of /pɛtʃ/. If the report of /h/ is indicative of the stop burst streaming, then a portion of the /p/ streamed. In addition, /tʃ/ was reported as the background percept, indicating C_2 streamed. We revisit the simultaneous streaming of both consonants in our discussion of Experiment 2.

When listeners reported hearing two streams, the perceptual objects in each stream were usually identified phonetically. However, in 20% of the streaming transformations, the object in the background stream was described as a repeating noise of some sort. This was true of all stimuli and occurred with slightly greater frequency in the intact than in the final and I + F conditions, suggesting that the nasal was more difficult to recognize as speech than the affricates.

The data presented thus far suggest that perceptual regrouping is one cause of verbal transformations and that the frequency and types of streaming reports are a function of the acoustic properties of the utterances. The ease with which the acoustic elements in a stream regroup should also be a function of the strength of alternative groupings (Bregman, 1978a). The more fragile they are, the less likely they will be heard, or heard for an extended period of time. Two analyses measuring this aspect of streaming were carried out across the three stimulus conditions.

Bregman (1978b) and Dorman et al. (1975) noted that stream formation can take time to build up, requiring many repetitions before the elements in a percept split and regroup to form separate streams. If the acoustic elements in the intact-condition stimuli were in fact more resistant to streaming than those in the final and I + F conditions, then regrouping should have occurred latest in the intact condition. By this same reasoning, the more resistant a percept is to streaming, the more stable the single percept should be, which should lead to longer stretches of hearing a single stream.

Evidence of both of these outcomes was found in the data. Listed in column 4 of Table 1 are the mean number of repetitions from trial onset (collapsed across participants and items) in the three conditions. It took almost twice as many repetitions to report a streaming transformation in the intact than in the final and I + F conditions, *F*(2, 24) = 12.42, *p* < .01. A similar pattern was found across conditions when the length of runs spent hearing a single stream was compared with the length of runs hearing two streams. These data, which do not include the data in column 4, are in the last two columns of Table 1. Looking first at the intact-condition data, one can see that the runs in which the stimuli were heard as forming two streams were slightly longer than those in which a single stream was heard, suggesting a mild bias to hear the utterance as a split percept. This bias was magnified two- and threefold in the final and I + F conditions. When compared with the means in the intact condition, the difference appears to be due more to a drop in one-stream reports than to an increase in two-stream reports, indicating that the final and I + F stimuli were much less stable when heard as a single stream than as two streams.

Statistical analyses reinforced these observations. A two-way analysis of variance (ANOVA) comparing the data across conditions (collapsed over participants) yielded a reliable main effect of number of streams, *F*(1, 12) = 18.67, *p* < .01, and an interaction of number of streams and stimulus condition, *F*(2, 24) = 5.52, *p* < .01. Comparisons within stimulus conditions showed that only the differences in the final and I + F conditions were reliable, *F*(1, 12) = 23.43, *p* < .01, and *F*(1, 12) = 5.34, *p* < .04, respectively. Comparison of the three one-stream reports showed that the drop from the intact to the final and I + F means was marginally reliable, *F*(2, 24) = 2.85, *p* < .08. The increase in two-stream reports in the final condition relative to the other two conditions reached significance, *F*(2, 24) = 14.24, *p* < .01.

Discussion

When listeners are instructed to pay attention to other characteristics of the auditory scene in addition to the verbal transformation, the reports that are provided suggest that the VTE can be caused by perceptual regrouping of the elements in the recycling stimulus. At least half of the time that listeners reported a change in their percept, more than one auditory stream was heard. The frequency and characteristics of these streaming transformations depended on the acoustic properties of the recycling stimulus, further suggesting that perceptual regrouping underlies some transformations. The intact-condition stimuli, with the strongest acoustic bonds between phonemes, were most resistant to regrouping, yielded the fewest streaming transformations, required the most repetitions before they split apart, and were heard as a single stream for only slightly fewer runs than they were heard as two streams. Nevertheless, these stimuli streamed almost 50% of the time, indicating that their acoustic elements can break off. This surprisingly high rate of streaming reports may be due in part to task demands. Participants were explicitly taught to listen for the presence of multiple streams, which may have inflated such reports (we return to this possibility when we discuss Experiment 2).

In contrast, the final and I + F stimuli behaved differently because they were created not only to split apart but to split apart in specific locations. C_1 cohered more tightly to the vowel than did C_2 in the final stimuli, resulting in only utterance-final streaming. C_1 was replaced in these same items with a less cohesive consonant in the I + F condition, which led to streaming in the initial location as well. The proclivity of these items to yield streaming transformations was also found in the immediacy with which the first streaming transformation was reported and the bias to hear long runs of the repeating stimulus as forming two streams.

The streaming transformations that listeners reported exhibited a considerable amount of regularity, with similar types of reports found across stimuli with the same phonemes. When /tʃ/ was heard in a separate stream, the transformation variants possessed similar characteristics (e.g., /pɛ/ and /wɛ/; /pɛt/ and /wɛt/). The same was true of /dʒ/ (e.g., /po/ and /lo/; /pod/ and /lod/). Note that some of these transformations included no trace of the affricate, giving the impression that it vanished when in fact it can be heard in the other stream. The streaming of /p/ in both /podʒ/ and /pɛtʃ/ yielded primarily /h/-initial transformations.

Reports of /h/ in place of /p/, or the complete deletion of C_1 or C_2, are not idiosyncratic to the stimuli of Experiment 1 but have been found in related studies. Clegg (1971) used CVs in which the vowel was held constant across stimuli and the consonant varied. Even at a relatively slow presentation rate (stimulus onset asynchrony estimated at 1.52 s), listeners occasionally reported that the initial consonant vanished. More than any other class of phonemes, /h/ was reported in place of voiceless stops (/p/, /t/, /k/). Goldstein and Lackner (1973) also found there was a tendency for listeners to report /h/ when the initial consonant of a CV was a voiceless stop. A few forms listed in Warren and Meyers (1987), in which the recycling stimulus was *to do*, were /h/-initial as well. Perceptual regrouping can readily explain these seemingly bizarre transformations.

Experiment 2

The results of Experiment 1 suggest that transformations due to streaming are the result of perceiving only a portion of the entire stimulus, namely, the foreground percept. The purpose of Experiment 2 was to replicate and build on this finding by using a less direct measure of regrouping but one that would identify the portion of the recycling stimulus that corresponds to the streaming transformation (i.e., foreground percept).

Instead of reporting transformations throughout a trial, listeners reported only the first transformation and then had to isolate the section of the stimulus that corresponded to it by moving two cursors, one at each end of the stimulus, to exclude the portion(s) of the signal that was not part of the transformation (i.e., the background percept). In essence, they edited a speech waveform (using only auditory cues), isolating the portion of the stimulus that formed what amounted to the foreground percept. The task resembles that used by Chalikia and Warren (1994), who had listeners adjust the width of a frequency filter to determine which portions of a recycling vowel sequence formed a separate spectral stream. Because Experiment 2 was designed to measure initial and final splitting of segments, the temporally based isolation technique seemed best suited for our purposes.

The stimuli of Experiment 1 were included in Experiment 2. If the transformations reported to those stimuli were the same in both experiments, then not only would our findings be replicated but the portion of the speech signal that corresponds to those transformations would be identified, pinpointing the structurally weak location in the utterances (i.e., where it splits into two streams). Within an utterance, differences in cursor placement should correspond to systematic differences in the transformations that are reported.

Method

Participants. Eighteen new undergraduates from the same pool as in Experiment 1 participated. One participant's data were discarded because the person did not follow instructions.

Stimuli. Because each trial was much shorter than in Experiment 1, 12 stimuli were used instead of only 6; all were CVC pseudowords. There were four stimulus conditions. In addition to the three conditions of Experiment 1, a fourth condition, initial, was added, in which C_1 was expected to stream off. There were three CVCs in each condition. In the intact, final, and I + F conditions, two of the CVCs were the ones used in Experiment 1, and a third one was added to each condition that was expected to behave similarly (/rʌl/, /rʌp/, and /tʌp/, respectively). The stimuli for the initial condition were /pɛm/, /pom/, and /tɛm/.

Stimuli were recorded and processed following the procedure described in Experiment 1, including new tokens of the six stimuli used in that experiment. Mean stimulus durations were very similar except in the final condition (intact = 462 ms; initial = 440 ms; final = 500 ms; I + F = 453 ms).

Procedure. Participants were tested one at a time. First, they were introduced to the VTE, and then they were instructed to listen for a transformation on each trial (no mention was made of auditory grouping). Once they heard a transformation, they were to report the transformation, which the experimenter transcribed, and then to isolate the transformation so that only it was audible, nothing else.

To isolate a transformation, participants pressed buttons on a response box that were linked to cursors (i.e., pointers) at the beginning and end of the stimulus sound file, controlling their movement inward into the file and outward toward the endpoints of the file. Only the portion of the sound file between the cursors was heard, and this section of the utterance continued to repeat while the transformation was isolated. Two buttons on the left side of the response box moved the cursor at the beginning of the sound file; one moved the cursor into the file, the other moved it back toward the beginning of the file. Two buttons on the right side of the response box performed the same function on the cursor at the end of the file. Each press moved the cursor 40 ms, a duration which we arrived at through our pilot work. The number of button presses from the beginning and end of the file were printed on the computer monitor in front of the listener to assist the participant in positioning the cursors. No other visual cues or information was provided. The next trial began after the participant signaled that the cursors were in their final positions.

In the practice session, two stimuli were used to familiarize listeners with the isolation technique. The 12 stimuli were presented in a randomly permuted order in the test session. Participants listened to the repeating pseudoword and then isolated the first transformation they heard by moving the left and right cursors. The experiment lasted 30 min.

Results

Data analyses first focused on cursor positioning in the waveforms and on comparing the transformations with those in Experiment 1. Waveforms of the stimuli, along with final mean cursor positions when the cursor was moved (measured in milliseconds from stimulus onset), are shown in Figure 1.[1] The stretch of speech that corresponded to the transformation was temporally longer than the remaining portion that was not isolated (314 ms vs. 145 ms), $t(16) = 7.64$, $p < .01$. Similar preferences for a temporally longer foreground percept have been reported in studies that used repeating vowel sequences (Chalikia & Warren, 1991) and tone sequences (Preusser, Garner, & Gottwald, 1970; Royer & Garner, 1970; Woodrow, 1951).

Looking first at the waveforms of the three CVCs in the intact condition, we found that the right cursor was moved for only /lom/ and /rʌl/, and then only into C_2. The left cursor was never moved. Listeners reported hearing only /lo/ when /lom/ was presented, which was one of the most frequent transformations in Experiment 1. Spectral analysis of /lom/ indicated that the right cursor was placed close to the start of nasalization, indicating that the transformation was based on the phonetic information contained in the isolated segment.

This coupling between the content of the isolated portion and the perceived transformation was not as tight with /rʌl/. Listeners reported transformations without /l/, but spectral analysis of the utterance indicated that the right cursor was positioned so that formant transitions into /l/ were still present in the isolated segment, though only weakly. This result may be common with liquids and not nasals, as it also was found with /pʌl/ in the initial condition for the 1 listener who isolated /po/. The 1 listener who heard /m/ stream from /pom/ completely removed all traces of the nasal. Because listeners never moved the cursors when presented with /wɛm/, /m/ does not appear to have split off and formed a separate stream. This outcome differs from Experiment 1 and may be due to the fact that the elements in the new recording of the pseudoword were more resistant to streaming.

Across the three initial-condition stimuli, the left cursor fell somewhere during the initial /p/ but always after the stop burst. For /pʌl/, it was at the beginning of aspiration. For /pom/, it was at the end of aspiration. For /pɛm/, it was even further into the utterance, occurring after the first pitch period of the vowel. Verbal reports across listeners and items fell into one of two categories, both of which were found in Experiment 1: Either the stop was completely deleted, in which case listeners reported a vowel-initial stimulus (e.g., /ɛm/), or listeners reported an /h/-initial utterance (e.g., /hɛm/). There were twice as many reports of the latter than of the former. These data replicate the finding in Experiment 1 that the acoustic elements of stops are much more likely to split off than those of liquids and glides. These latter segments are not immune from streaming, and inspection of the final cursor placements in Experiment 2 suggests that formant transitions are the break point.

In the final condition, the right cursor was usually placed between the vocalic segment and the final consonant for all three items. With /lodʒ/, listeners reported hearing either /lo/ or /lod/, the two most frequent transformations reported in Experiment 1. Nine of 13 participants heard the formant transitions specifying /d/, and their final cursor positions reflected this, being 52 ms closer to frication than those of participants who heard only /lo/. A similar outcome occurred with /wɛtʃ/. Of the 12 listeners for whom /tʃ/ streamed, 11 reported /wɛt/ and 1 reported /wɛ/. Both of these forms were reported in Experiment 1. As with /lodʒ/, formant transitions out of the vowel contributed to the perception of /t/. Inclusion of a small fragment of /tʃ/ in the isolated segment may have been heard as a release burst, contributing to the perception of the stop. With /rʌp/, the stop burst streamed from the remainder of the utterance. The 4 listeners for whom /r/ split off instead of /p/ all reported /ʌp/ and placed the cursor at the end of the third formant (F3) transition.

In the I + F condition, the outcomes of the initial and final conditions were replicated. Placement of the left cursor fell somewhere in the aspirated portion of the stop consonant across all three stimuli. As in the initial condition, listeners reported either vowel-initial or /h/-initial transformations only. Placement of the right cursors was in approximately the same locations as in the final condition. Listeners' reports paralleled those in the final condition as well. For /podʒ/, there were more reports containing a final /d/ (/pod/) than the vowel only (/po/). For /pɛtʃ/, all listeners reported a final /t/. With /tʌp/, the /p/ streamed off and nothing was heard in its place. As a few reports in Experiment 1 hinted, unique to the I + F condition were transformations in which both the initial and final segments streamed off simultaneously. Such occurrences were fairly rare, constituting only 18% of the reports in this condition, and conformed to the types of responses just described (e.g., /hɛt/ given /pɛtʃ/; /hod/ given /podʒ/).

Because listeners never reported the number of streams they heard, there is no direct evidence that they in fact heard two. However, the similarity of the reports across Experiments 1 and 2 strongly suggests that they did. Additional evidence to bolster this claim was found when we compared reports across the four stimulus conditions by using some of the same measures of streaming as in Experiment 1. Because many of the transformations that listeners reported in Experiment 2 corresponded to streaming transformations in Experiment 1, reports in Experiment 2 were operationally defined as streaming transformations if at least one of the cursors was moved inward into the utterance, removing a portion of the signal, *and* the consonant at that word position (i.e., initial or final) was not reported or /h/ was heard in place of a stop. By this definition, 61% of the transformations were due to perceptual regrouping.

Column 1 of Table 2 shows the proportion of streaming transformations in each condition. Streaming transformations were least frequent in the intact condition and occurred three times as often in the other three conditions. A chi-square test comparing the frequency of streaming transformations across conditions was reliable, $\chi^2(3, N = 124) = 16.91$, $p < .01$. Further comparisons

[1] If a nonstreaming transformation was reported, listeners still attempted to isolate the corresponding segment, which resulted in one of two outcomes: Either the participant gave up trying or the cursors were moved minimally or not at all.

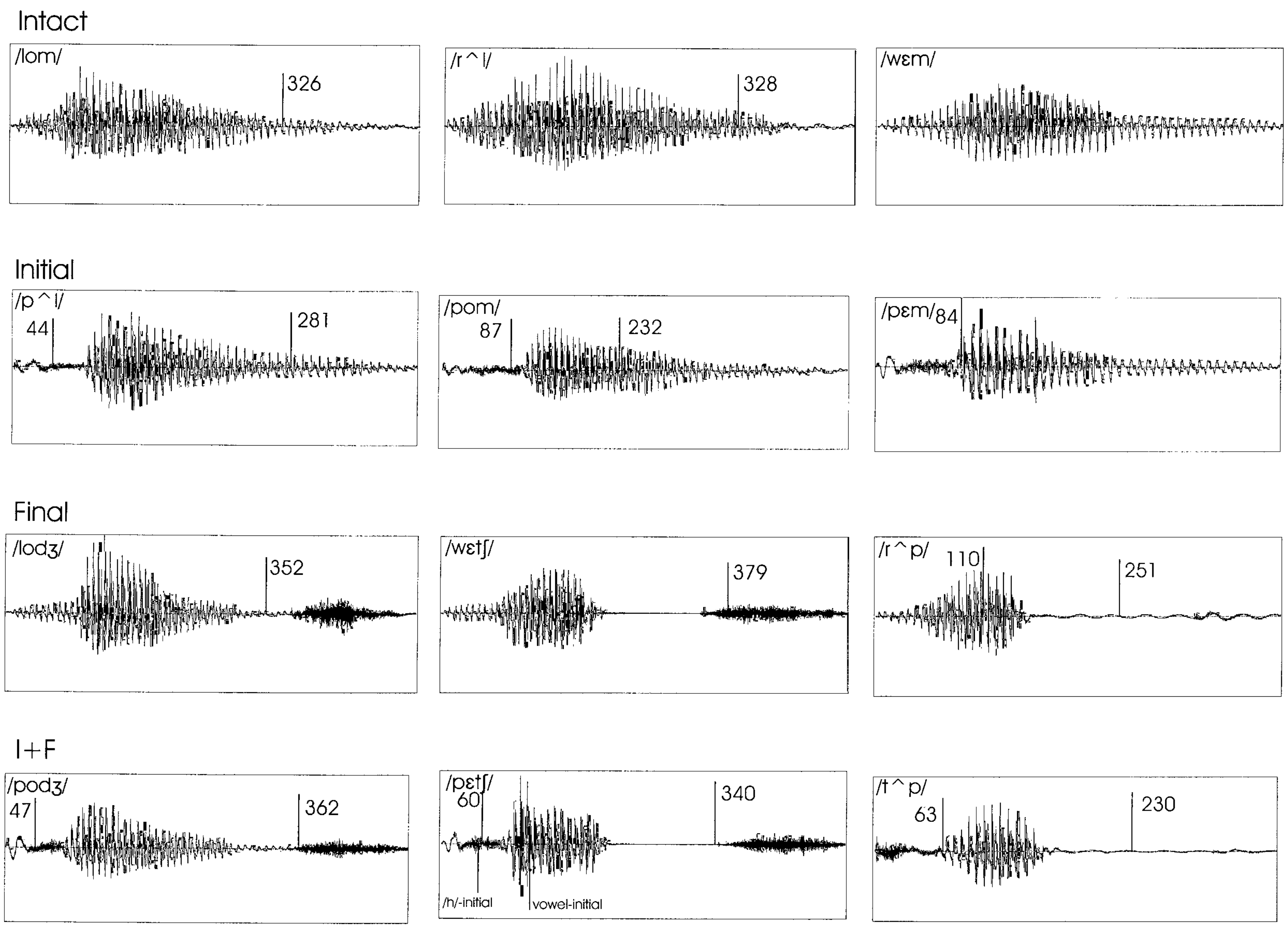

Figure 1. Waveforms of the 12 stimuli used in Experiment 2. Vertical lines represent mean cursor positions measured in milliseconds from stimulus onset. I + F = initial plus final.

Table 2
Transformation Measures in the Three Stimulus Conditions of Experiment 2

	Proportion of streaming transformations		
Condition	Overall	Initial location	Final location
Experiment 2			
Intact	.22	.00	1.00
Initial	.63	.94	.06
Final	.78	.10	.90
I + F	.80	.61	.39
Replication of Experiment 2			
Intact	.25	.36	.64
Initial	.60	1.00	.00
Final	.75	.00	1.00
I + F	.77	.53	.47

Note. I + F = initial plus final.

showed that the intact condition differed significantly from the other three conditions, which did not differ among themselves. Comparison of these means with their counterparts in Experiment 1 shows that the values in Experiment 2 span a much wider range, a finding that is due primarily to the drop in the intact condition. This difference may be a result of only one (the first) transformation being recorded on each trial in Experiment 2. It may also be further evidence that the high rate of streaming transformations reported in the intact condition in Experiment 1 is partially due to demand characteristics from instructing participants to listen for multiple streams.

Columns 2 and 3 of Table 2 contain the same data broken down by whether the initial or final consonant was not included in the segment that was isolated (i.e., it formed the background percept). As in Experiment 1, the manipulation of stimulus cohesiveness had the intended effect. In the initial condition, the initial segment streamed virtually all of the time. An effect of almost equal magnitude but in the opposite direction was obtained in the final condition. In the I + F condition, both the initial and final segments streamed. Although there was a greater tendency for the initial segment to stream, this finding was not significant, χ^2 (1, $N = 41$) = 2.47, $p > .10$. The means in the I + F condition in Experiment 1 pattern in the opposite direction, which may also be due to differences in methodology between the experiments. In the intact condition, the final segment always split off in the few cases in which a streaming transformation was heard.

Discussion

Listeners in Experiment 2 were not explicitly made aware of the phenomenon of streaming, yet their data provide compelling evidence of perceptual regrouping. Many of the streaming transformations reported in Experiment 1 were also reported by listeners in Experiment 2, suggesting that the stimuli split into separate streams. This interpretation is reinforced by the cursor placement data, which indicate cursors were moved to eliminate the part of the utterance that corresponded to the background streams in Experiment 1: affricates and stop bursts. Finally, when the number of streaming transformations was tallied across the four stimulus conditions, the data resembled those of Experiment 1. Consonants that were acoustically least similar to the vowel split off far more often than those that were acoustically similar and did so in the expected utterance locations. The consistency and convergence of the findings across both experiments provides strong evidence that perceptual regrouping is one cause of verbal transformations.

The data of Experiment 2 also revealed the close correspondence between the streaming transformation that was reported and the phonetic information specified in the isolated segment. With the exception of the two utterances containing final /l/, the transformation that was heard was determined in large part by the phonetic information present in the isolated segment. One of the most compelling examples of this, which highlights listeners' sensitivity to the phonetic contents of the isolated segment, was found with the /p/-initial utterances. When streaming occurred, the stop burst always split off, with all listeners moving the cursor inward past the burst. This was a much rarer occurrence with the following aspirated portion. When aspiration was included in the isolated segment (83% of the time), listeners always reported hearing /h/. Collapsed over stimuli and listeners, mean cursor position for reports of this type was 56 ms from stimulus onset. When aspiration fell outside the isolated segment, listeners never reported /h/ but instead heard a vowel-initial transformation. Mean cursor position was 107 ms, almost twice as far into the utterance. To understand how different these cursor positions are, they are overlaid on the waveform /pEtʃ/ in Figure 1. Note that the vowel-initial cursor was moved past the aspiration and well into the vowel. A similar bimodal positioning of the left cursor was found for /tʌp/.

The consistency of listeners' reports to these voiceless stops and also to the utterance-final affricates suggests that there are two reliable markers of streaming transformations: One is deletion of a consonant at a word boundary; the other is replacement of an initial voiceless stop with /h/.

Finally, a replication of Experiment 2 was carried out to demonstrate that the results were not specific to the stimuli. Sixteen listeners were tested on 16 stimuli (11 new), which were more variable in structure, including CVs, CCVs, and CCVCs, some of which were words used in past studies on the VTE (e.g., *see, ripe, tress, truce*; Warren, 1961b). The streaming transformation measures are shown in the bottom half of Table 2 and are very similar to those in the top half. The proportion of streaming reports again differed across stimulus conditions, $\chi^2(3, N = 155) = 18.73$, $p < .01$, and all stimuli, including the words, split apart as anticipated. For example, /s/ in *see* and /p/ in *ripe* streamed off, as did /t/ and /s/ in *truce* and *tress*.[2]

Experiment 3

The purpose of Experiment 3 was to examine whether the perceptual process underlying streaming transformations is specific to speech. On the one hand, the acoustic elements of speech appear to stream for some of the same reasons tones and other nonspeech objects stream: Spectral (e.g., frequency, timbre)

[2] Figures containing the waveforms and the mean cursor positions for the stimuli used in the replication of Experiment 2 can be found at http://1pl.psy.ohio-state.edu. Example stimuli are also provided there.

changes over a given time interval are too great, destroying cohesion among sequentially occurring elements and leading to one perceptual grouping being abandoned in favor of another (Bregman, 1978a; Chalikia & Warren, 1994; Cole & Scott, 1973; Van Noorden, 1975; Warren et al., 1969; Wessel, 1979). These parallels suggest that the process that causes perceptual regrouping is the same for both classes of stimuli.

On the other hand, the VTE is very much a verbal illusion. Listeners hear phonetic percepts, indicating the stimulus is processed linguistically. Remez, Rubin, Berns, Pardo, and Lang (1994) made a persuasive case that the principles of perceptual organization that influence how sequences of tones and other nonspeech sounds group together (e.g., auditory scene analysis; Bregman, 1990) are at present not sufficient to explain why the acoustics of speech cohere to form a single perceptual event (however, see Barker & Cooke, 1999; Rosenthal & Okuno, 1998). In particular, the alternation of consonants (many of which are aperiodic noises) and vowels (periodic stretches) in speech production creates marked discontinuities in frequency and frequency trajectories that should cause the speech stream to crumble into countless streams, yet it rarely does. Thus, streaming transformations reported in the preceding experiments might be due to perceptual processes specific to speech, processes that would not be invoked if those same stimuli were heard as nonspeech.

To address this issue, we compared the foreground percepts that listeners isolated when sinewave analogs of speech were heard as recycling speech and as recycling nonspeech. Acoustically, sinewave speech consists primarily of three time-varying sinusoids that track the center frequency of the first three formants of speech (Remez, Rubin, Pisoni, & Carrell, 1981). The phonetic quality of sinewave speech is sufficiently poor so that, when first encountered, it is generally heard as speech only if the listener is informed it is speech. Otherwise, listeners report hearing whistles or computer sounds. However, just as concealed figures in visual displays are impossible not to see once one notices them (e.g., the outline of a Dalmatian or a cow's face hidden in a picture of black dots on a white background; see Held, 1974, and Street, 1931, for other examples), once the perceiver knows the stimulus is speech, it is difficult not to hear it as speech.

Sinewave speech is an ideal stimulus for a test of processing specificity because differences in performance cannot be attributed to differences in stimuli, only to differences in how the stimuli were perceived. Because differences have been found as a function of how sinewave speech is processed (e.g., Best, Morrongiello, & Robson, 1981; Best, Studdert-Kennedy, Manuel, & Rubin-Spitz, 1989; Remez et al., 1994), it is reasonable to expect that differences will also be found if the process responsible for perceptual regrouping is different for speech and nonspeech.

Method

Participants. Forty-seven new undergraduates from the same pool as in Experiment 1 were recruited. Ten of them participated in the sinewave-speech condition. Of the 37 participants who were tested in the tone condition, the data of 27 were not analyzed because these listeners heard at least one sinewave analog as speech at some point during the experiment.[3]

Stimuli. Four sinewave analogs of words beginning with "/s/+stop" clusters served as the test stimuli. We chose /s/+stop clusters because a sinewave token of /s/ is clear and identifiable in the signal. Also, transformations reported in Shoaf and Pitt (in press) suggest /s/ streams before stops. Two words, *steady* and *screen*, were excised from sinewave sentences originally recorded by Remez and Rubin (Remez et al., 1981). The other two, *sketch* and *spell*, were created using natural tokens spoken by Mark A. Pitt as templates. Three additional sinewave tokens were created and used in filler trials. They were chosen such that the portion that we believed would be most likely to stream occurred word-finally and was not always a fricative. These items were included to force participants to listen carefully to the transformations instead of automatically responding to all stimuli in the same way (e.g., isolating the final portion). Two sinewave practice items were also created.

Procedure. The procedure was identical to that of Experiment 2 except for modifications in the instructions. In the sinewave-speech condition, participants were told that they would hear speech that was produced by a computer. They first engaged in a short sinewave-speech familiarization session to ensure that the sinewave stimuli would be clearly heard as speech. Participants in the tone condition received the same instructions as those in the sinewave-speech condition except that the sinewave stimuli were not referred to as speech but as computer sounds and the like. No familiarization with the sinewave stimuli was provided. In addition, preliminary testing in the tone condition revealed that many listeners began hearing the stimuli phonetically at some point during the test session. To lessen this possibility, we digitally reversed the two sinewave practice stimuli and the three filler stimuli so that they would not be heard as spoken utterances. This necessitated splicing and recombining segments of the fillers so that the portion that was assumed to stream would still occur in the intended stimulus location (i.e., at the end of the stimulus).

In the practice session, two naturally spoken words were used to familiarize listeners with hearing streaming transformations and moving the cursor to isolate the transformation. Two additional practice trials with sinewave tokens were then presented, with listeners in the sinewave-speech condition receiving speech instructions and listeners in the tone condition receiving nonspeech instructions. In the test session, seven test stimuli were presented in two counterbalanced orders, with a filler stimulus always occurring between two test stimuli. In the sinewave-speech condition, listeners reported the transformation before isolating it. Because it was sometimes difficult for listeners in the tone condition to describe the transformation immediately, they were allowed to isolate the foreground sound prior to describing it, which gave them more time to provide an accurate description of what was heard. Participants were tested one at a time in a 45-min session.

Results

Except for 2 listeners who did not move the cursors on one trial, all listeners isolated a portion of each of the utterances, suggesting that the stimuli split into two perceptual streams. In the sinewave-speech condition, the isolated section always included the vocalic portion of the utterance and was temporally longer than the remaining segment, which always included /s/ (363 vs. 179 ms, respectively), $t(9) = 6.21$, $p < .01$. A weaker trend was found in the tone condition (258 vs. 195 ms), $t(9) = 1.67$, $p < .13$, but there were reversals, with listeners isolating the shorter segment 15% of the time (always the portion corresponding to /s/). Because the stimuli were the same in the two instruction conditions, the reversal was probably due to the relative meaningfulness of the two

[3] Although listeners never heard the sinewave analog in the tone condition as speech when it was first presented, repetitive presentation of the stimuli caused some listeners to hear them phonetically as the actual word or one similar to it at some point during the experiment. High participant-replacement rates in experiments that have used sinewave speech have been reported (Sawusch & Gagnon, 1995). The problem was particularly acute in the current experiment because the phonetic quality of the sinewave tokens was good, and many sinewave stimuli were used.

segments. When perceived as speech, the vocalic portion is much more phonetically salient than the /s/ frication. This difference in saliency may not have been as great when the stimuli were heard as nonspeech sounds. Because the most informative analysis between the two conditions would be one in which cursor positions are compared with the same portion of the stimulus that was perceived as the foreground, the data from these six trials were excluded from the analysis. Their inclusion increases variability in the data but does not change interpretation of the outcomes in any meaningful way.

Shown in Figure 2 are waveforms of the four stimuli along with mean beginning and ending cursor positions in the sinewave-speech and tone conditions. Comparison of the cursor positions across waveforms reveals a number of commonalities in the sinewave-speech and tone conditions that indicate the stimuli streamed in very similar ways when heard as speech and nonspeech. To begin with, across all four stimuli the left cursor was always placed in the silent (closure) interval between the fricative and the onset of voicing, except for *steady*, in which it was placed immediately after the formant transitions into the vowel in the tone condition. With the exception of the few cases mentioned in the preceding paragraph, /s/ was always heard in the background in both the sinewave-speech and tone conditions.

The isolated portion of the word was not simply the remainder of the utterance. The right cursor was positioned well into the stimulus for most words. With *screen*, the cursor was placed prior

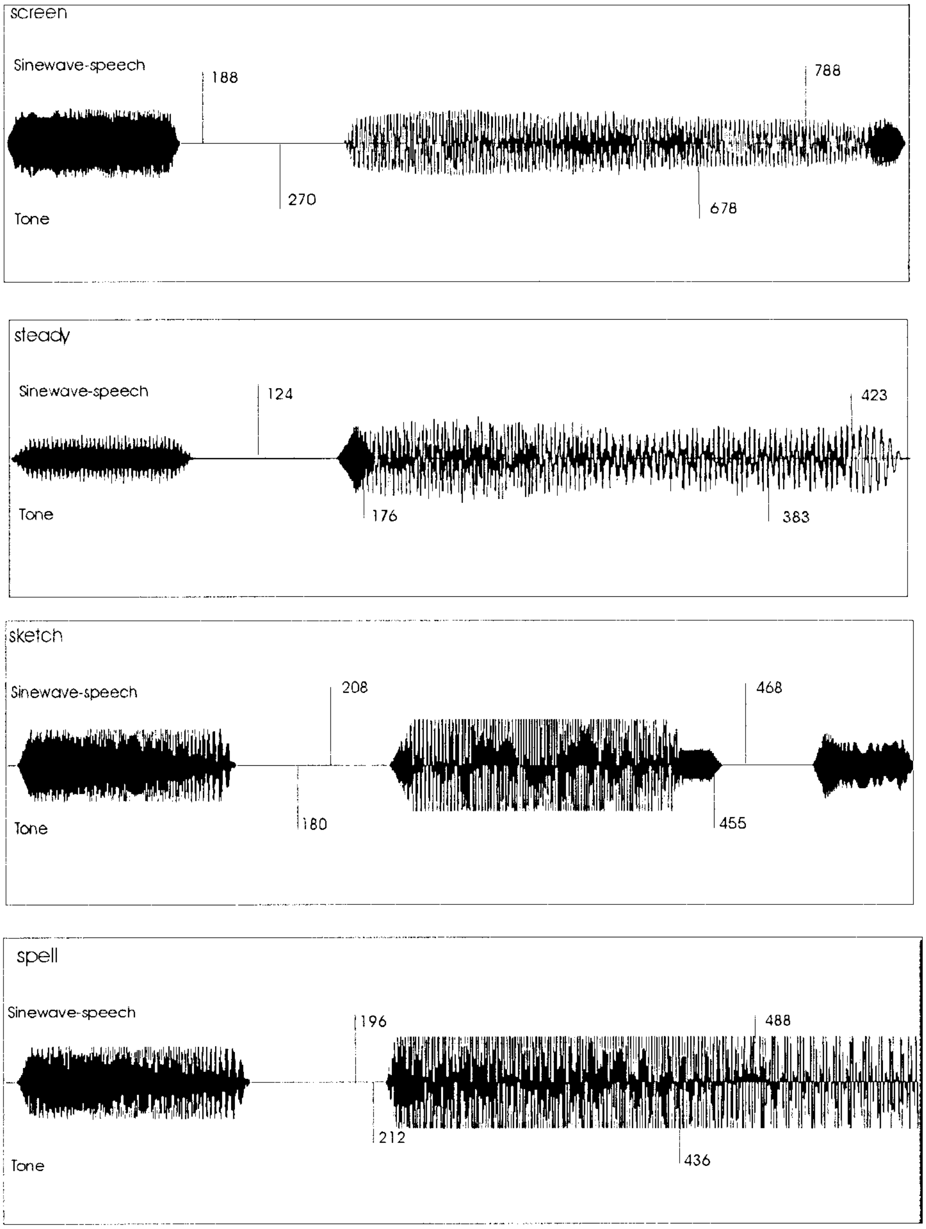

Figure 2. Waveforms of the four sinewave stimuli used in Experiment 3, along with mean cursor positions in the sinewave-speech and tone conditions.

to the portion of the signal that corresponds to the nasal resonances in the tone condition. In the sinewave-speech condition, nasal resonances were included in the isolated segment. With *steady*, the cursor was positioned in the final vowel. With *sketch*, the final affricate streamed off in the sinewave and tone conditions. With *spell*, the cursor was placed at or slightly after the third formant transition out of the vowel and into /l/.

These results demonstrate that the isolated segments across stimuli contained approximately the same acoustic information in the sinewave and tone conditions, with placement of the right cursor being the primary cause of the differences between conditions. A statistical comparison between conditions on the location of the right cursor was marginally reliable, $t(18) = 2.00$, $p < .06$, with cursor placement being further toward the center of the stimulus when it was heard as nonspeech. No reliable difference was obtained for the left cursor. Although for some stimuli mean cursor position differed by more than 50 ms across the sinewave-speech and tone conditions, statistical comparisons on individual items failed to reach significance in all cases except the left cursor position of *sketch*, $t(18) = 2.88$, $p < .01$. Differences in placement of the right cursor led to the isolated segment being longer in duration when heard as speech than as nonspeech, $t(18) = 3.18$, $p < .01$. This difference was also reliable for three of the four items when analyzed individually ($p < .04$ or lower in all cases), with *sketch* being the only stimulus for which there was a reversal of this pattern, $t(18) = 1.80$, $p < .09$.

Differences in cursor placement are likely to be due to differences in what listeners were trying to isolate in the two conditions and to listeners' sensitivity to changes in the percept when the stimulus was heard as speech versus nonspeech. When they heard the stimulus as speech, listeners would have positioned the cursors to include all of the phonemes in the transformation that was perceived, which required moving the cursors far enough toward the stimulus endpoints so as to include the beginning and ending phonemes of the transformation. To do this successfully, a listener would have to detect subtle changes in the phonetic quality of the stimulus, such as the presence of nasalization or a liquid. When listeners heard the stimulus as nonspeech, such changes are not likely to have been as meaningful, and because of this, they were probably less influential in cursor positioning.

Evidence to support the preceding account can be found in the variability in cursor placement across conditions. The standard deviation of the position of the right cursor, averaged over stimuli, was substantially larger in the tone than the sinewave-speech conditions (*SD* = 100 vs. 65 ms, respectively). The difference was in the same direction for the left cursor, although its magnitude was smaller (*SD* = 73 vs. 64 ms, respectively). The consistency in cursor placement across listeners in the sinewave-speech condition translated into consistency in verbal reports as well. For each stimulus, most reports were identical or were slight variations from each other: /rin/, /rim/, and /grim/ constituted most of the reports for *screen*; /ɛdi/ was reported by 8 listeners when *steady* was presented; /gɛt/, /kɛ/, and /ɛtʃ/ were reported for *sketch*; and finally, /ɛl/ and /ɪl/ were the most frequent responses to *spell*. In the tone condition, the transformations were described as various pulsating and animal sounds, and changes along auditory dimensions such as pitch and rhythm.

The sinewave-speech and tone conditions also differed in the number of repetitions required to form a stream. When heard as speech as compared with nonspeech, the sinewave analogs were listened to for more than twice as many repetitions *before* the cursors were moved (40 vs. 17, respectively), $t(18) = 3.50$, $p < .01$). Not only was this true for listeners but it held across all four stimuli, suggesting that the elements of the sinewave analogs cohered more tightly as a single stream when heard as speech than as nonspeech.

Discussion

The similarity of the isolation data across the two instruction conditions suggests that some of the same principles of perceptual organization operate when processing speech as well as nonspeech. Unfortunately, this outcome makes it difficult to determine the specificity of the perceptual process that underlies streaming transformations, as the results are equally amenable to the two theoretical viewpoints. On the one hand, the data could reflect the operation of a single, general-purpose process that governs the perceptual reorganization of speech and nonspeech. On the other hand, they could be accounted for by separate speech and nonspeech grouping processes, each with similar performance characteristics (Kuhl, 1987).

The data in the sinewave-speech condition also replicated the findings of Experiment 2. The acoustic elements at word boundaries, most notably those corresponding to sibilants, split off from the remainder and formed a background stream, and the verbal transformation corresponded to the temporally longer stretch of the signal. The data from the current experiment also identified a few other properties of verbal, as opposed to nonverbal, transformations. First, identification of the isolation points appears to be governed in part by the phonetic content of the transformation. Second, when heard as speech, sinewave analogs were more resistant to streaming than when they were heard as nonspeech.

General Discussion

The VTE is a fascinating and simultaneously perplexing illusion. Identification of its cause is intended not only to unmask the mystery of the illusion itself but also to contribute to our understanding of the processes underlying perception. The approach taken in this study was to learn what causes verbal transformations by analyzing the forms that listeners reported in experimental setups designed to reveal the operation of a hypothesized process.

Experiments 1–3 were tests of the proposal that perceptual regrouping is one such process. Experiment 1 demonstrated that listeners hear multiple streams, with the verbal transformation corresponding to the foreground percept and the unreported segment corresponding to the background percept. Supporting evidence for the regrouping account was obtained in Experiment 2 and its replication, in which listeners not only reported many of the same transformations as in Experiment 1 but also identified the acoustic chunks of the utterances that corresponded to the transformations. Experiment 3 showed that whether the stimulus is heard as speech or as nonspeech has an influence on grouping.

One characteristic of streaming transformations that emerges from the isolation data of Experiment 2 is that what is heard is determined by the acoustic information present in the foreground stream. This tight link is most evident with the /p/-initial and affricate-final stimuli. As mentioned previously, when the stop burst plus aspiration was eliminated, transformations began with a vowel (e.g., /ɛtʃ/, /odʒ/), but when some of the aspiration was part of the foreground, /h/-initial transformations were reported. Similarly, whether or not a final consonant was reported depended on how far the end cursor was moved inward. The consistency of such responding across stimuli and across listeners underscores the priority of bottom-up input in determining the phonetic content of streaming transformations.

The findings of Experiment 3 suggest that regrouping is not solely bottom-up. When sinewave analogs were heard as speech rather than nonspeech, transformations tended to be longer and took twice as long to form. One interpretation of this finding is that it provides evidence of the involvement of top-down processes in the perceptual organization of speech. Such influences have been demonstrated in the perceptual organization of tones and melodies. For example, listeners who are given the name of a tune before hearing it are much more accurate in identifying it when distractor notes are woven into the melody (Dowling, 1992; Dowling, Lung, & Herrbold, 1987). Something similar may be going on when listeners are informed that the sinewave replicas are speech. For example, knowledge of how words are spoken could influence perceptual organization so that the resulting foreground stream (i.e., the transformation) is linguistically plausible and does not contain uninterpretable chunks.

Because the sinewave replicas were heard as words as opposed to pseudowords when participants were given speech instructions, the differences between the tone and sinewave-speech conditions may have been due to lexical influences on perceptual grouping. That is, lexical memory might affect the strength with which the auditory elements of a word cohere. This idea is supported by the fact that listeners in the sinewave-speech condition took twice as long to begin isolating the transformation. Additional evidence to support this proposal came from a recent study comparing verbal transformations to words and pseudowords. Shoaf and Pitt (in press) found that words transformed back into the veridical percept more often than did pseudowords, suggesting that the lexical representation of the word directly influenced the perceptual organization of speech. Nygaard (1993) obtained similar results, finding that lexical status affected the perceptual fusion of dichotically presented fragments of an utterance. Together, these findings suggest that lexical processes may counter the effects of repetitive presentation that lead to streaming and help bind the acoustic elements of a word together. They also make the point that word perception involves more than processing the phonetic segments that make up a word; it also includes processes that assist in binding the elements of those words together.

In some ways, it should come as no surprise that perceptual regrouping is one cause of the VTE. The experimental setup is ripe for it according to all we know about what contributes to the streaming of speech and other auditory objects. Large, rapid alterations in frequency lead to streaming (van Noorden, 1975). High-frequency frication will split off from a following vowel and form a separate stream when the CV is presented repetitively (Cole & Scott, 1973). Repetitive presentation of an utterance promotes alternative groupings of the acoustic elements because the elements occur at regular, predictable points in time (Bregman, 1990). Finally, the phenomenological cousin of verbal transformations, phonemic transformations are due in part to streaming by frequency. Chalikia and Warren (1994) had listeners adjust the width of a bandpass filter to define the frequency region occupied by the verbal forms that participants reported hearing while listening to a repeating sequence of six vowels. Their results demonstrated that listeners heard two streams, one above and one below approximately 1500 Hz. Some principles of auditory grouping may apply uniformly to all verbal sequences regardless of how they are constructed or presented.

Although all three experiments were designed to elicit streaming transformations, listeners' reports sometimes varied in other ways as well. As mentioned in Experiment 1, most of these differences consisted of vowel and consonant substitutions and insertions, yielding variants of the veridical stimulus. The consistency and frequency of these transformations (12% of responses in Experiment 1) suggests they may reflect the operation of one or more additional perceptual processes. A host of other factors have also been shown to influence listeners' reports, such as the phonological complexity of the stimulus and its meaning (Warren, 1961b). Clearly, perceptual regrouping is only one of potentially many causes of verbal transformations. Perhaps further systematic analysis of listeners' transformations will be useful in identifying additional causes.

References

Barker, J., & Cooke, M. (1999). Is the sine-wave speech cocktail party worth attending? *Speech Communication, 27,* 159–174.

Best, C. J., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics, 29,* 191–211.

Best, C. T., Studdert-Kennedy, M., Manuel, S., & Rubin-Spitz, J. (1989). Discovering phonetic coherence in acoustic patterns. *Perception & Psychophysics, 45,* 237–250.

Bregman, A. S. (1978a). Auditory streaming: Competition among alternative organizations. *Perception & Psychophysics, 23,* 391–398.

Bregman, A. S. (1978b). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 380–387.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: MIT Press.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology, 89,* 244–249.

Brochard, R., Drake, C., Botte, M., & McAdams, S. (1999). Perceptual organization of complex auditory sequences: Effect of number of simultaneous subsequences and frequency separation. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1742–1759.

Chalikia, M. H., & Warren, R. M. (1991). Phonemic transformations: Mapping the illusory organization of steady-state vowel sequences. *Language and Speech, 34,* 109–143.

Chalikia, M. H., & Warren, R. M. (1994). Spectral fissioning in phonemic transformations. *Perception & Psychophysics, 55,* 218–226.

Clegg, J. M. (1971). Verbal transformations on repeated listening to some English consonants. *British Journal of Psychology, 62,* 303–309.

Cole, R. A., & Scott, B. (1973). Perception of temporal order in speech: The role of vowel transitions. *Canadian Journal of Psychology, 27,* 441–449.

Cooper, A. M., Whalen, D. H., & Fowler, C. A. (1986). P-centers are unaffected by phonetic categorization. *Perception & Psychophysics, 39,* 187–196.

Cullinan, W. L., Erdos, E., Schaefer, R., & Tekieli, M. E. (1977). Perception of temporal order of vowels and consonant–vowel syllables. *Journal of Speech and Hearing Research, 20,* 742–751.

de Jong, K. J. (1994). The correlation of p-center adjustments with articulatory and acoustic events. *Perception & Psychophysics, 56,* 447–460.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27,* 769–773.

Diehl, R. L., Kluender, K. R., & Parker, E. M. (1985). Are selective adaptation and contrast effects really distinct? *Journal of Experimental Psychology: Human Perception and Performance, 11,* 209–220.

Diehl, R. L., Parker, E. M., & Kluender, K. R. (1985). On the pitfalls of "ptolemaic" psychology: A reply to Sawusch and Mullennix. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 251–256.

Ditzinger, T., Tuller, B., & Kelso, J. A. S. (1997). Temporal patterning in an auditory illusion: The verbal transformation effect. *Biological Cybernetics, 77,* 23–30.

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 and Performance, 1,* 121–129.

Dowling, W. J. (1992). Perceptual grouping attention and expectancy in listening to music. In J. Sundberg (Ed.), *Gluing tones: Grouping in music composition, performance and listening* (Vol. 72, pp. 77–98). Stockholm, Sweden: Royal Swedish Academy of Music.

Dowling, W. J., Lung, K. M.-T., & Herrbold, S. (1987). Aiming attention in pitch and time in the perception of interleaved melodies. *Perception & Psychophysics, 41,* 642–656.

Goldstein, L. M., & Lackner, J. R. (1973). Alterations of the phonetic coding of speech sounds during repetition. *Cognition, 2,* 279–297.

Held, R. (1974). *Image, object, and illusion.* San Francisco: Freeman.

Kuhl, P. K. (1987). The special-mechanisms debate in speech research: Categorization tests on animals and infants. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 355–386). New York: Cambridge University Press.

Lackner, J. R., & Goldstein, L. M. (1974). Primary auditory stream segregation of repeated consonant–vowel sequences. *Journal of the Acoustical Society of America, 56,* 1651–1652.

MacKay, D. G., Wulf, G., Yin, C., & Abrams, L. (1993). Relations between word perception and production: New theory and data on the verbal transformation effect. *Journal of Memory and Language, 32,* 624–646.

Natsoulas, T. (1965). A study of the verbal-transformation effect. *American Journal of Psychology, 78,* 257–263.

Nygaard, L. (1993). Phonetic coherence in duplex perception: Effects of acoustic differences and lexical status. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 268–286.

Preusser, D., Garner, W. R., & Gottwald, R. L. (1970). Perceptual organization of two-element temporal patterns as a function of their component one-element patterns. *American Journal of Psychology, 83,* 151–170.

Reisberg, D., Smith, D., Baxter, D., & Sonenshine, M. (1989). Enacted auditory images are ambiguous; pure auditory images are not. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 4*(A), 619–641.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review, 101,* 129–156.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981, May 22). Speech perception without traditional speech cues. *Science, 212,* 947–950.

Rosenthal, D. F., & Okuno, H. G. (1998). *Computational auditory scene analysis.* Mahwah, NJ: Erlbaum.

Royer, F. L., & Garner, W. R. (1970). Perceptual organization of nine-element auditory temporal patterns. *Perception & Psychophysics, 7,* 115–120.

Sawusch, J. R., & Gagnon, D. A. (1995). Auditory coding, cues, and coherence in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 635–652.

Sawusch, J. R., & Jusczyk, P. (1981). Adaptation and contrast in the perception of voicing. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 408–421.

Sawusch, J. R., & Mullennix, J. W. (1985). When selective adaptation and contrast effects are distinct: A reply to Diehl, Kluender, and Parker. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 242–250.

Shoaf, L. S., & Pitt, M. A. (in press). *A test of node stability in node structure theory. Perception & Psychophysics.*

Street, R. F. (1931). *A Gestalt completion test: A study of a cross section of intellect.* New York: Columbia University, Bureau of Publication.

van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences.* Unpublished doctoral dissertation, Eindhoven University of Technology, Eindhoven, the Netherlands.

Warren, R. M. (1961a). Illusory changes in repeated words: Differences between young adults and the aged. *American Journal of Psychology, 74,* 506–516.

Warren, R. M. (1961b). Illusory changes of distinct speech upon repetition: The verbal transformation effect. *British Journal of Psychology, 52,* 249–258.

Warren, R. M. (1968). Verbal transformation effect and auditory perceptual mechanisms. *Psychological Bulletin, 70,* 261–270.

Warren, R. M. (1976). Auditory illusions and perceptual processes. In N. Lass (Ed.), *Contemporary Issues in Experimental Phonetics* (pp. 389–416). New York: Academic Press.

Warren, R. M. (1985). Criterion shift rule and perceptual homeostasis. *Psychological Review, 92,* 574–584.

Warren, R. M., Healy, E. W., & Chalikia, M. H. (1996). The vowel-sequence illusion: Intrasubject stability and intersubject agreement of syllabic forms. *Journal of the Acoustical Society of America, 100,* 2452–2461.

Warren, R. M., & Meyers, M. D. (1987). Effects of listening to repeated syllables: Category boundary shifts versus verbal transformations. *Journal of Phonetics, 15,* 169–181.

Warren, R. M, Obusek, C. J., Farmer, R. M., & Warren, R. P. (1969, May 2). Auditory sequence: Confusion of patterns other than speech or music. *Science, 164,* 586–587.

Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer Music Journal, 3,* 45–52.

Woodrow, H. (1951). Time perception. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology* (pp. 1224–1236). New York: Wiley.

Received May 3, 2000
Revision received January 29, 2001
Accepted May 11, 2001 ■