Auditory Illusions and Confusions

These failures of perception are studied because they isolate and clarify some fundamental processes that normally lead to accuracy of perception and appropriate interpretation of ambiguous sounds

by Richard M. Warren and Roslyn P. Warren

For more than a century visual illusions have been of particular interest to students of perception. Although they are in effect misjudgments of the real world, they apparently reflect the operation of fundamental perceptual mechanisms, and they serve to isolate and clarify visual processes that are normally inaccessible to investigation. Auditory illusions, on the other hand, have received little scientific attention. Until recently the fleeting nature of auditory stimuli made it difficult to create, control and reproduce sound patterns as readily as visual ones. The tape recorder made it easy to manipulate sounds, and yet for a time there was little examination of auditory illusions, perhaps because there was no historical tradition to build on—no puzzles inherited from the experimental psychologists of the past century, as there were in the case of optical illusions. Some new investigations, however, have led to the discovery of illusions in hearing that help to explain the human ability to extract information from fleeting patterns of sound. These investigations have also led to the identification of confusions in hearing that help to explain some limitations of that ability.

Consider for a moment that you are at a convention banquet. While you are still finishing your dinner the after-dinner speeches begin. The clatter of dishes masks some of the speech sounds, as do occasional coughs from your neighbors and your own munching. Nonetheless, you may be able to understand what the speaker is saying by utilizing the information that reaches you during intervals that are relatively free of these interfering noises. In order to understand how speech perception functions in the presence of transient noises, we and Charles J. Obusek did some experiments last year in our laboratory at the University of Wisconsin at Milwaukee. First we recorded the sentence "The state governors met with their respective legislatures convening in the capital city." Then we carefully cut out of the tape recording of the sentence one phoneme, or speech sound: the first "s" in "legislatures." We also cut out enough of the preceding and following phonemes to remove any transitional cues to the identity of the missing speech sound. Finally, we spliced the recorded sound of a cough of the same duration into the tape to replace the deleted segment.

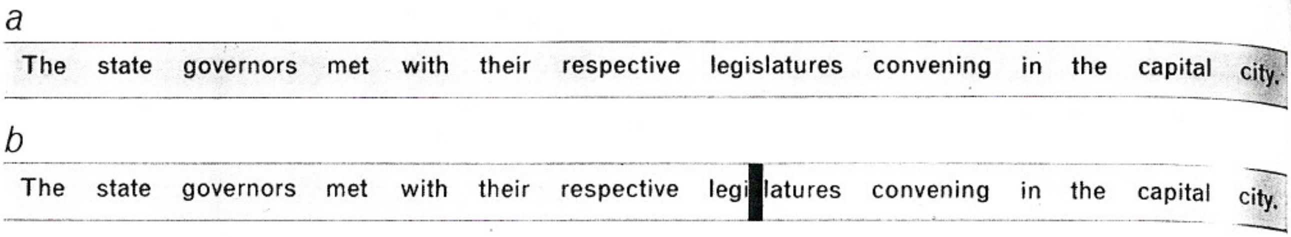
When this doctored sentence was played to listeners, we found that we had created an extremely compelling illusion: the missing speech sound was heard as clearly as were any of the phonemes that were physically present. We called this phenomenon "phonemic restoration." Even on hearing the sentence again, after having been told that a sound was missing, our subjects could not distinguish the illusory sound from the real one. One might expect that the missing phoneme could be identified by locating the position of the cough, but this strategy was of no help. The cough had no clear location in the sentence; it seemed to coexist with other speech sounds without interfering with their intelligibility. Phonemic restoration also occurred with other sounds, such as a buzz or tone, when these sounds were as loud as or louder than the loudest sound in the sentence. Moreover, phonemic restorations were not limited to single speech sounds. The entire syllable "gis" in "legislatures" was heard clearly when it was replaced by an extraneous sound of the same duration.

We did find a condition in which the missing sound was not restored. When a silent gap replaced the "s" in "legislatures," the gap could be located within the sentence and the missing sound identified. In visual terms, it was as if an erasure of a letter in a printed text could be detected, whereas an opaque blot over the same symbol would result in illusory perception of the obliterated letter, with the blot appearing as a transparent smear over another portion of the text [*see top illustration on pages 32 and 33*]. Of course, in vision a blot can be localized readily, and even the more elusive "proofreader's illusions" can be eliminated when the reader is told in advance just where the error in the text occurs. With phonemic restorations, however, knowledge of the nature of the extraneous sound and of the identity of the missing phoneme does not prevent clear perception of the missing sound—even when the stimulus is played to the listener as many times as he wishes.

The inability to localize an extraneous sound in a sentence was first reported in 1960 by the British workers Peter Ladefoged and Donald E. Broadbent. Since they employed brief intrusive sounds (clicks and short hisses) and took care that no phoneme was obliterated, phonemic restorations did not arise. Similar short, nonmasking extraneous sounds were later used by a group at the Massachusetts Institute of Technology that included Jerry A. Fodor, Merrill F. Garrett and Thomas Bever. They have reported that systematic errors in locating the clicks are caused by various features of sentence structure, and they have used the errors to explore those features.

Perceptual synthesis of the phoneme is accomplished on the basis of verbal context. In the case of the missing "s" in "legislatures" the context prior to the absent sound suffices for identification. What about a sentence so constructed that the context necessary to identify an obliterated sound does not come until

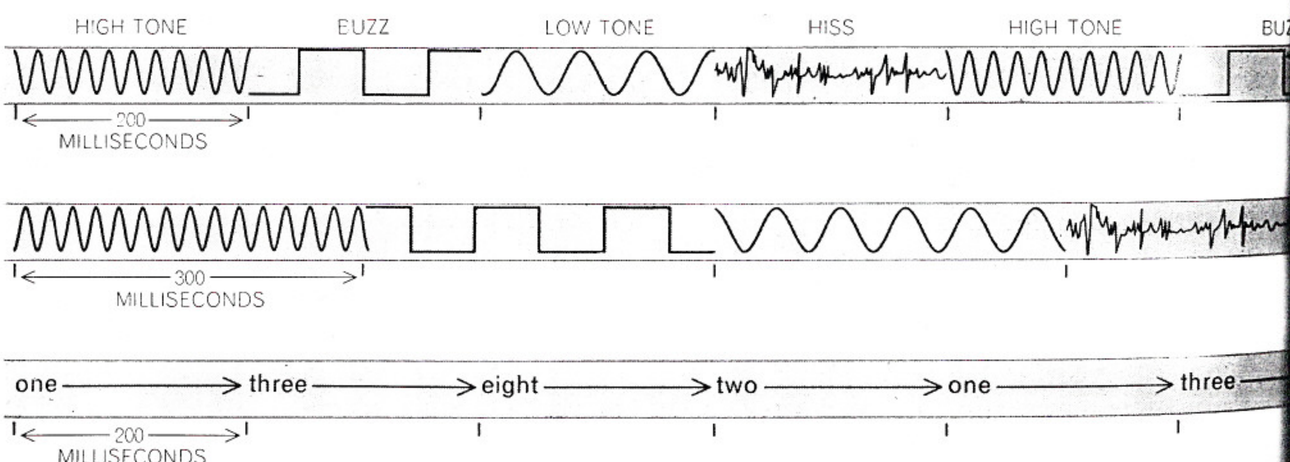
UDITORY ILLUSIONS are investigated in the authors' laboratory. The subject, listening through headphones to a stimulus signal generated by the equipment in the background and reproduced by the tape recorder, reports to the experimenter on what he hears.



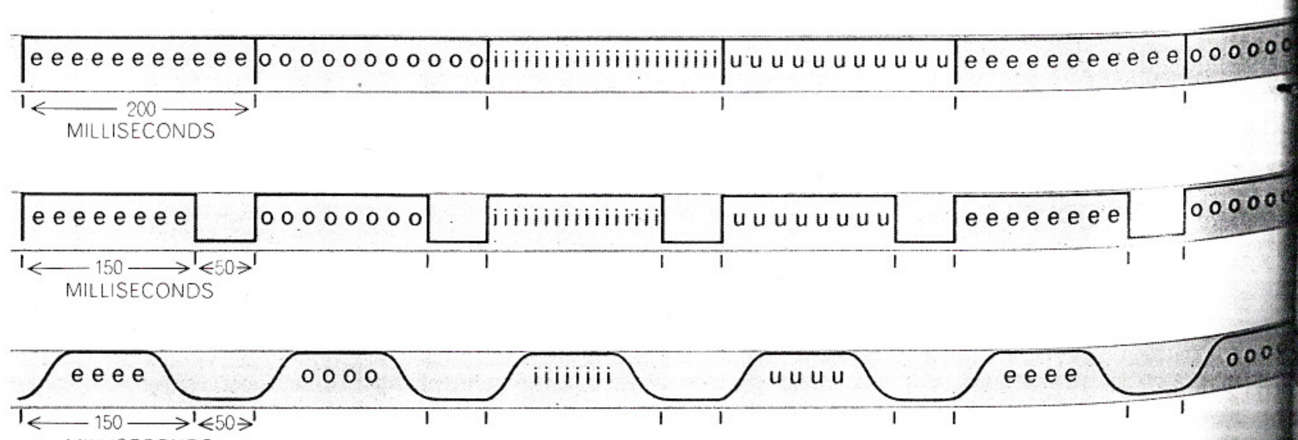
PHONEMIC RESTORATION is an illusion that shows the importance of context in determining what sound is heard. A sentence was recorded on tape (*a*). Then the first "s" in "legislatures" was excised and a cough of the same duration (*black rectangle*) was spliced in its place (*b*). When the altered sentence was played to subjects, the missing "s" was heard clearly (*c*) and localization of

later? With the symbol * representing a loud cough that replaces a speech sound, consider a spoken sentence beginning, "It was found that the *eel was on the ___." The context provided by the last word in the sentence should resolve the ambiguity and determine the appropriate phonemic restoration. Among the words that could complete the sentence are "axle," "shoe," "orange" and "table." Each implies a different speech sound for the preceding word fragment, respectively "wheel," "heel," "peel" and "meal." Preliminary studies by Gary Sherman in our laboratory have indicated that the listener does experience the appropriate phonemic restoration, apparently by storing the incomplete information until the necessary context is supplied so that the required phoneme can be synthesized. We are still investigating the influence of such factors as the duration of extraneous sounds in relation to the duration of the missing phoneme and the maximum temporal separation between the ambiguous word fragment and the resolving context that will still permit phonemic restoration.

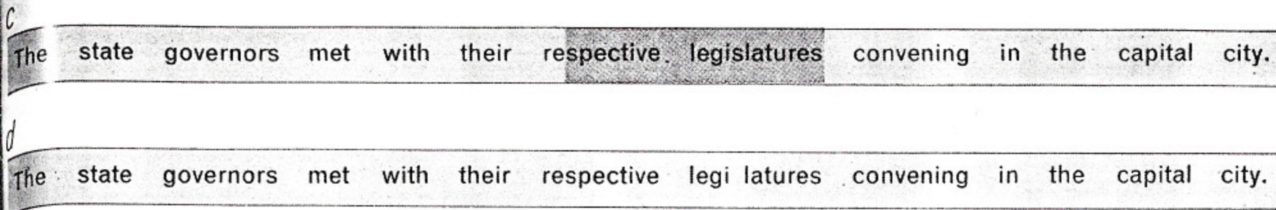
The use of subsequent context for correcting errors had been suggested on logical grounds by George A. Miller of Rockefeller University. He reasoned that unless some such strategy were available, a mistake once made while listening to spoken discourse would cause errors in interpreting the following portions of the message to pile up, until the entire system eventually stalled. The long delays in muscular activity that have been observed in the skilled transcription of an incoming message also suggest that storage of incoming lan-



TEMPORAL CONFUSION was observed when a high tone, a buzz, a low tone and a hiss (represented here schematically), each lasting 200 milliseconds, were presented repeatedly (*top*). Subjects could not report the sequence of the sounds properly whether they tried



FOUR VOWEL SOUNDS were used in another experiment on temporal confusion. When the vowel sounds of "beet," "boot," "bit" and "but" were presented at a sustained level for 200 milliseconds their sequence could not be determined (*top*). Deleting 50 milli

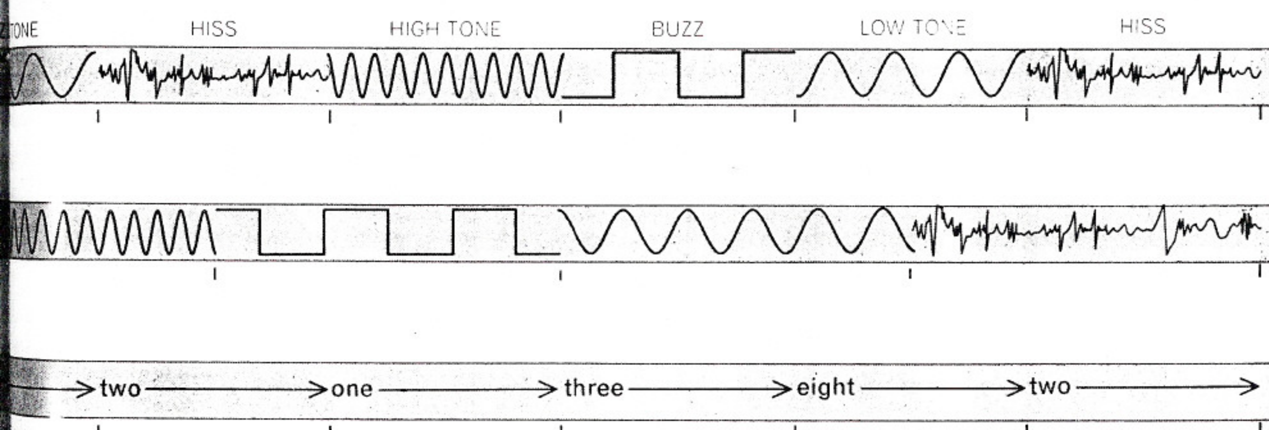


the cough was indefinite; when required to guess the location, subjects generally missed the correct position by several phonemes, as indicated (*gray area*). When a silent gap, rather than a cough, replaced the "s," the gap could be located and the missing sound could be identified (*d*). This illustration, like those that follow, is necessarily an approximate representation of an auditory effect.

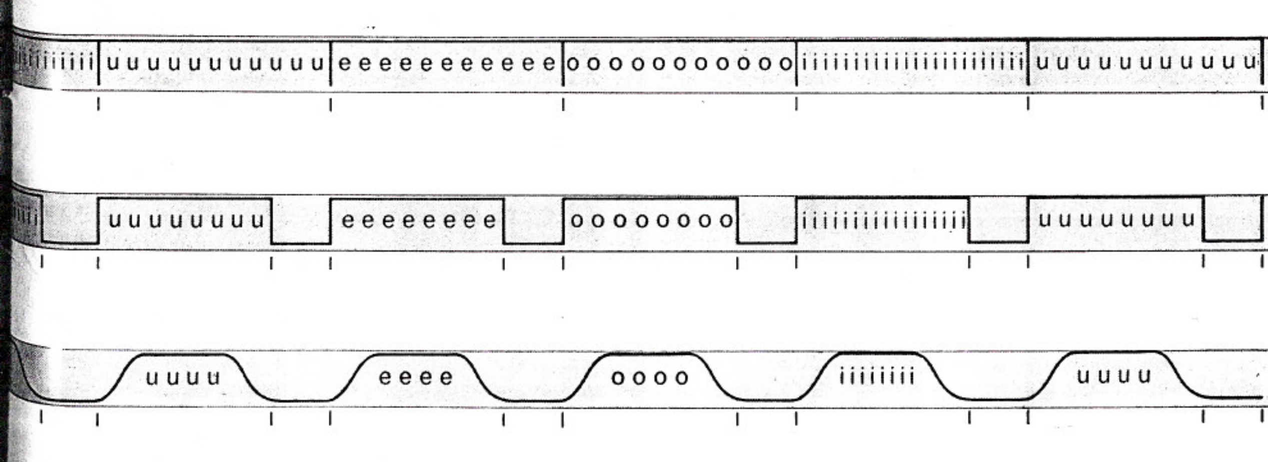
guage information is associated with error correction. In the 1890's William Bryan and Noble Harter noted that highly skilled telegraphers listening to Morse code did not transcribe the auditory signals that constituted a word until some six to 12 words after the signals were heard. If subsequent portions of the message could not provide helpful context, as in the case of stock quotations or transmissions in cipher, the telegraphers changed their strategy and followed the message much more closely in time. Telegraph companies charged higher rates for sending such messages precisely because they lacked redundant context, were therefore much more difficult to receive and had to be transmitted more slowly.

This telegrapher's technique illustrates a surprising relation that one encounters again and again in perception: The development of an extremely complex procedure for data processing is necessary to achieve the deceptive impression of an "easy" perceptual task. From time to time other workers have noted the delay between language input and motor response. In 1925 William Book observed the similarity between typewriting and code transcription, reporting that in the case of an expert typist "attention was pushed ahead of the hands as far as possible (usually four or five words)."

Inability to locate the position of extraneous sounds in sentences represents a failure in the detection of temporal order. It might be thought that this temporal confusion results from a conflict between verbal and nonverbal



o do so verbally or by ordering four cards, each representing a sound. When sounds lasted 300 milliseconds, subjects could order them with cards (*middle*). When spoken digits were substituted for sounds, it was easy for subjects to report their order (*bottom*).



conds of each sound and replacing it with silence (*middle*) allowed half of the subjects to determine the sequence. The sequence was readily determined when vowels were given normal qualities of gradual onset and decay, suggested by curves (*bottom*).

a

tresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstres

b

tress tress tress tress tress tress tress tress stress stress stress stress stress stress stress stress stress dress dre

dress dress dress dress floris floris floris floris floris florist florist florist florist florist Joyce Joyce Joyce Joyce Joyce

VERBAL TRANSFORMATION EFFECT is noted when subjects listen to a distinct recording of a single word repeated on a loop of tape (*a*). One might expect a kind of reversal effect, with the stimulus "tress" perceived sometimes as "rest." Instead lack of context

modes of perception. Recent observations in our laboratory have indicated, however, that inability to detect sequence is not restricted to verbal-nonverbal interactions. In 1968, during an experiment on loudness, we noted to our surprise that listeners could not tell the order of three successive sounds repeated over and over without pauses. The sounds—a hiss, a tone and a buzz—each lasted a fifth of a second (200 milliseconds) and were recorded on a tape that was then spliced to form a loop. The duration of each sound was quite long compared with the 70- to 80-millisecond average for a phoneme in speech and was well within the temporal range used in music for the successive notes of melodies; the hiss, tone and buzz could each be heard clearly. Yet it was impossible to tell the order of the sounds. The pattern swirled by, the temporal structure tantalizingly just beyond one's grasp.

It might be thought that a little advance planning would make the task easy. It should be possible, for example, to concentrate on one of the sounds (say the hiss) and then decide whether the sound that follows it is a tone or a buzz; this single decision would fix the third sound in the remaining slot and solve the problem. In practice, however, the single decision cannot be made with accuracy. Out of 50 listeners we found that only 22 named the order correctly—slightly fewer than the 50 percent correct answers that would be expected by chance alone.

This seemed at first to contradict the findings of earlier studies. Results reported by Ira Hirsh of the St. Louis Central Institute for the Deaf and by others had indicated that temporal resolution of such sounds as tones, hisses and buzzes should be possible down to a separation of about 20 milliseconds—even less time than is required for accurate temporal ordering of the sounds forming speech or music. These values, however, had been based on pairs of sounds. The subjects listened to a single pair (such as a tone and a hiss) and reported their order. It was possible, we reasoned, that subjects could say which sound came first and which last not by actually perceiving the temporal order as such but by detecting which of the sounds occurred at either the onset or the termination of the stimulus pair. In 1959 Broadbent and Ladefoged had suggested that the ability of their subjects to order pairs of sounds might be based on the "quality" of the pair as a whole. Could that "quality" be determined by which sound was present at the onset and/or the termination of the brief pair?

With threshold judgments of this kind, when subjects are working at the limit of their ability, introspection as to how they make their decisions is particularly difficult; they simply cannot say. To determine what criteria are actually being used one must rely on experiments. We returned to our recycled 200-millisecond stimuli (hiss, tone, buzz) that could not be ordered, but this time we inserted a three-second interval of silence between successive presentations of the three-sound sequence. Most listeners could now identify the first and last items in the series correctly, somewhat more accurately in the case of the last sound. This supported our suspicion that "sequence perception" with pairs of sounds represents a special case and is really perception of onset and termination.

In order to examine further the perception of temporal order in the absence of onset and termination cues, we employed a variety of repeated four-item sequences. The chance of guessing the order correctly, starting with whichever sound one chooses, is one in six. With a sequence consisting of a high tone (a frequency of 1,000 hertz, or 1,000 cycles per second), a buzz (40-hertz square wave), a low tone (796 hertz) and a hiss (2,000-hertz octave band of noise), each lasting 200 milliseconds, correct responses were only at the level of chance. It was necessary to increase the duration of each item to between about 300 and 700 milliseconds (the exact value depending on practice and the response procedure) to obtain a correct identification of sequence from half of the subjects tested. For durations of 300 milliseconds or more, calling out the order of the sounds resulted in more errors than arranging four cards, each bearing the name of one sound, in the appropriate sequence.

We noticed a curious feature of the four-item sequences: listeners frequently could not tell at first how many different sounds were present in the series. The apparent disappearance of one or sometimes even two items could be minimized by telling the listener the number of sounds there were and by first introducing each sound alone. This illusory absence of stimuli could not account completely for the inability to perceive sequence, however: even people who heard the four sounds clearly could not report their sequence. We also found that repetition was not in itself a barrier to sequence perception. When four spoken digits, each lasting 200 milliseconds, were recorded separately (to avoid transitional cues), spliced into a loop and repeated over and over, the subjects perceived the order at once and with certainty.

This great difference between the temporal perception of verbal and of nonverbal stimuli suggested that we could use perception of sequence in an effort to establish which attributes of sounds are responsible for speechlike characteristics. We cut four 200-millisecond segments out of extended statements of separate vowels held at a fixed level for several seconds. When these tape segments were spliced into a loop and played back, the listener heard a repeated sequence of four steady vowels following one another without pauses. Since no speaker can possibly change from one vowel to another in this way, without a transition or a pause, the sequence sounded curiously artificial, like tresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstresstress

ess dress dress dress dress dress tress tress tress tress tress tress Joyce Joyce Joyce Joyce Joyce Joyce dress

ss dress dress stress stress stress stress stress dress dress dress dress dress purse purse purse purse purse . . .

has a more profound effect: most subjects experience illusory changes involving substantial distortion of the stimulus. A man listening to "tress" repeated 360 times in three minutes heard 16 changes involving eight different words, some illustrated here (*b*).

crude attempts to synthesize speech sounds electronically.

Our subjects did no better than chance the first time they attempted to judge the order of the sounds. By deleting a 50-millisecond portion of each sustained vowel and replacing it with a silent gap, we made the sequence sound more like normal speech, and then identification of order was possible for half of a new group of subjects. The subjects approached a perfect score only when we presented vowels of the same duration (150 milliseconds separated by 50 milliseconds of silence) but recorded with the normal qualities of vocal onset and decay that are characteristic of separate short utterances of vowel sounds. It appears, in short, that accurate perception of temporal order may be possible only for sequences that resemble those encountered in speech and in music—special sequences in which the component sounds are linked together, following specific rules, into coherent passages.

During the 1950's Colin Cherry of the Imperial College of Science and Technology in London wrote about the "cocktail-party problem," the task of attending to one chosen conversation among several equally audible conversations. Apparently such cues as voice quality and spatial localization help the listener to keep fixed on a single voice among many. When a person attends to one of these verbal sequences, he excludes the others, so that presumably it would not be possible for him to relate the temporal position of a phoneme in one conversation (or other extraneous sounds such as coughs) to the temporal position of phonemes in the attended conversation. Such observations lead us to speculate that the inability to perceive the correct order of stimuli that do not form integrated sequences of speech or music may not represent a flaw or defect of our perceptual skills. Rather, this restriction of temporal pattern perception may be an essential step in the continual process of extracting intelligible signals from the ubiquitous background of noise.

Musical and verbal passages have an organization based on the temporal order of their sounds; this organization furnishes a context for the individual sounds. Verbal context, as we pointed out above, can determine completely the synthesis of illusory speech sounds; phonemic restorations are heard when the context is clear but part of the stimulus is absent. Another illusion arises when the stimulus is clear but the context is absent. If one listens to a clear recording of a word or phrase repeated over and over, having only itself as context, illusory changes occur in what the voice seems to be saying. Any word or phrase is subject to these illusory changes, usually with considerable phonetic distortion and frequently with semantic linkages. These illusory words are heard quite clearly, and listeners find it difficult to believe they are hearing a single auditory pattern repeated on a loop of tape. As an example of the kind of changes heard, a subject listening to "tress" repeated without pause heard distinctly, within the course of a few minutes, such illusory forms as "dress," "stress," "Joyce," "floris," "florist" and "purse." This illusion, which we call the verbal transformation effect, has provided unexpected glimpses of hitherto unexplored perceptual mechanisms for organizing speech sounds into words and sentences.

The implications of the verbal transformation illusion were not appreciated fully in 1958, when one of us (Richard Warren) and Richard L. Gregory first reported the discovery of "an auditory analogue of the visual reversible figure." We had been looking for an auditory illusion resembling the one observed in such ambiguous figures as the Necker cube, whose faces seem to pop into different perspective orientations as one looks at it. We reasoned that ambiguous auditory patterns would undergo similar illusory shifts; for example, the word "rest" repeated clearly over and over without pause should shift to "tress," then back to "rest" and so on. We did find such closed-loop shifts but we also found some other illusory changes—to "dress" and "Esther," for instance. At the time, although we noted that perceptual distortion of the stimulus had occurred, we considered it only a curious side effect.

Further study by the present authors has drawn attention to basic differences between the visual and auditory illusions, however. The auditory effect is not limited to ambiguous patterns; any word or phrase will do. Changes are impossible to predict, vary greatly from individual to individual and often involve considerable distortion of the stimulus pattern. A subject listening to the word "see" repeated over and over may hear a phrase as far removed from the stimulus as "lunchtime," particularly if the time is about noon! Changes occur frequently: when a single word is repeated twice a second for three minutes, the average young adult hears about 30 changes involving about six different forms.

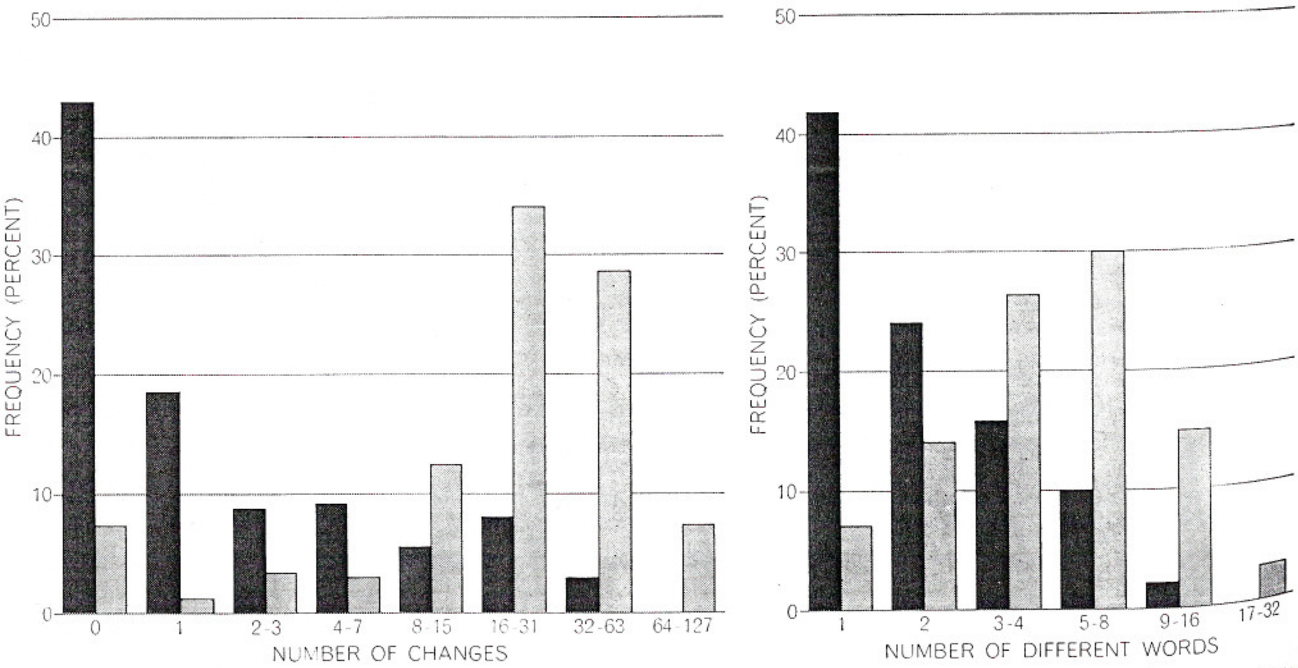
There are some remarkable effects of age on the frequency of verbal transformations and the types of illusory changes. These age differences seem to reflect basic changes in the way in which a person processes verbal input over a life-span. Children at the age of five experience either very few or no verbal transformations. At six half the children tested heard illusory changes, and those who did experienced them at the rapid rate characteristic of older children. By the age of eight all the children tested heard verbal transformations. The rate of illusory changes apparently remains approximately constant into the twenties and then declines slowly during the middle years; for listeners over 65 the rate was found to be only a fifth the rate for young adults and was approximately equal to the rate for five-year-olds. This decrease after middle age is not due directly to any decrease in auditory acuity with aging. Actually the aged are generally more accurate in this task than the young, reporting common English stimulus words correctly and continuing to respond to the stimulus as it actually is—the same word repeated over and over without change. Moreover, if young adults hear a word played indistinctly against a background of noise (which should simulate a decrease in acuity), they still hear many more illusory changes than the aged.

Besides counting the number of changes, we have examined the groupings of speech sounds to determine the units of perceptual organization at different ages. Children respond in terms of the sounds of English but may group them in ways not found in the language. For example, with the word "tress" repeated over and over, a child might report "sreb" even though the initial "sr" sequence is not found in English words. Young adults group speech sounds only in ways that are permitted in English, but they do report nonsense syllables: given the stimulus "tress," they might report "tresh" as one of the sounds they hear. Older people, on the other hand, report only meaningful words. Presented with "tress," they tend to hear "tress" continuously, and when infrequent changes do occur, they usually are to such closely related forms as "dress." If an older person is presented with a repeated nonsense syllable, there is an interesting result. If "flime" is the stimulus, for example, the older listener generally distorts the word into a phonetically close English word such as "slime" and tends to stay with the sense-making (but illusory) word throughout.

Our observations with verbal transformations have suggested that as people grow older they employ different perceptual mechanisms appropriate to their familiarity with language and their functional capacities, both of which change with age. We believe specific mechanisms associated with the skilled use of verbal context underlie the age differences in the frequency and nature of verbal transformations. Repeated words do not flow past us as normal components in the stream of language do; like a vortex, they move without progressing. In the absence of the semantic and grammatical confirmation ordinarily provided by verbal context, perception of repeated words becomes unstable for all but the very young and the old. And since each successive perceptual organization is subject to the same lack of stabilizing context, it suffers the fate of its predecessor.

The absence of illusory changes at age five suggests that young children have not yet reached the stage in language development where storage with skilled reorganization comes into play. The loss of susceptibility in alert and healthy elderly listeners suggests that they no longer have the functional capacity for this mechanism. It is rather well established that short-term memory is less effective in the aged when intervening activity is required between input and retrieval. Concurrent processes of coding, storing, comparing and reorganizing may therefore not be possible, so that the optimum strategy is to employ only the past context of the message as an aid to organization of the current input. The fact that in the presence of repeated stimuli the aged report only meaningful words is consistent with this view. If this interpretation is correct, one would expect that phonemic restoration for elderly people would be limited to replacement of speech sounds identified by prior context; the use of subsequent context, in the manner of young adults, would not be possible. We plan to do experiments testing this prediction.

In summary, it appears that phonemic restorations and verbal transformations provide new techniques for studying the perceptual organization of heard speech, particularly the grouping of speech sounds, the correction of the listener's errors and the resolution of acoustic ambiguities. The observations we have described for the perception of auditory sequence indicate that special perceptual treatment of the sounds of speech (and music) allow us to extract order and meaning from what would otherwise be a world of auditory chaos. It is curious that in studying illusions and confusions we encounter mechanisms that ensure accurate perception and the appropriate interpretation of ambiguities.



AGE DIFFERENCES in the frequency of verbal transformations are shown for two of the age groups tested. The bars indicate the number of changes from one form to another (*left*) and the number of different forms (*right*) perceived during three-minute tests by subjects 18 to 25 (*gray*) and 62 to 86 years old (*black*). Differences reflect changes in the perceptual processing of speech.