338

Reprint Series
22 May 1981, Volume 212, pp. 947–950

SCIENCE

Speech Perception Without Traditional Speech Cues

Robert E. Remez, Philip E. Rubin, David B. Pisoni, and Thomas D. Carrell

Speech Perception Without Traditional Speech Cues

Abstract. *A three-tone sinusoidal replica of a naturally produced utterance was identified by listeners, despite the readily apparent unnatural speech quality of the signal. The time-varying properties of these highly artificial acoustic signals are apparently sufficient to support perception of the linguistic message in the absence of traditional acoustic cues for phonetic segments.*

A person listening to a continuously changing natural speech signal perceives a sequence of linguistic elements. Researchers have attempted to characterize this perceptual process by analyzing the acoustic properties of speech signals that specify the linguistic content (*1*). In the present study, however, listeners perceived linguistic significance in acoustic patterns with properties differing substantially from those traditionally held to underlie speech perception. And, although listeners accurately reported the linguistic content of these acoustic patterns, the signal was also perceived, simultaneously, not to be speech. These novel findings imply that the process of speech perception makes use of time-varying acoustic properties that are more abstract than the spectra and speech cues typically studied in speech research (*1*).

The stimuli used in our study consisted of time-varying sinusoidal patterns that followed the changing formant center frequencies (the natural resonances of the supralaryngeal vocal tract) of a naturally produced utterance. The sentence "Where were you a year ago?" was spoken by an adult male, digitized at the rate of 10 kHz, and analyzed in sampled-data format. Frequency and amplitude values were derived every 15 msec for the center frequencies of the first three formants by the method of linear predictive coding (LPC) (*2*). These values were hand-smoothed in some portions to ensure continuity and were used as synthesis parameters for a digital sine wave synthesizer. Three time-varying sinusoids were then generated to match the LPC-derived center frequencies and amplitudes of the first three formants, respectively, of the natural utterance. Figure 1 shows narrowband and wideband spectrograms of the original utterance and a narrowband spectrogram of its replica formed by the three time-varying sinusoids.

Although our synthetic stimuli were designed to preserve the frequency and amplitude variation of natural speech formants, the three-tone patterns differ from natural speech in several prominent ways. First, the energy spectra of the tones differ greatly from those of natural and synthetic speech. Voiced speech sounds, produced by pulsed laryngeal excitation of the supralaryngeal cavities, exhibit a characteristic spectrum of harmonically related values (*3*, *4*). Because the frequencies of the individual tones in our stimuli follow the formant center frequencies, the components of the spectrum at any moment are not necessarily related as harmonics of a common fundamental. In essence, the three-tone pattern does not consist of harmonic spectra, although natural voiced speech does.

Second, the short-time spectra of the tone stimuli lack the broadband formant structure that is also characteristic of speech (including whispered speech). Because the resonant properties of the supralaryngeal vocal tract introduce short-time amplitude maxima and minima across the harmonic spectrum of energy generated at the larynx, some frequency regions contain harmonics with more energy than neighboring regions (*5*). Our tone stimuli consist of no more than three sinusoids, so no energy is present in the spectrum except at the particular frequencies of each tone. Thus, the short-time spectra of the tone stimuli are also distinct in this way from the energy spectra of natural speech. There is no formant structure to the three-tone complexes, although the tones do exhibit acoustic energy at frequencies identical to the center frequencies of the formants of the original, natural utterance.

Third, the dynamic spectral properties of speech and tone stimuli are quite different. Across phonetic segments, the relative energy of each of the harmonics of the speech spectrum changes. Formant center frequencies may be computed by following the changes in amplitude maxima of the harmonic spectrum. However, natural speech signals do not exhibit continuous variation in formant frequency. Rather, laryngeal activity in voiced speech creates distinct pulses characterized by a formant structure. Thus, changes in formant structure, particularly when observed in wideband spectrograms, may erroneously appear to contain continuous formant variation over time. Figure 1B displays a wideband spectrogram in which the fine-grained amplitude differences are averaged over frequency to derive the formant pattern. In contrast to the case in speech, each tone in our stimuli continuously follows the computed peak of a changing resonance of the natural utter-

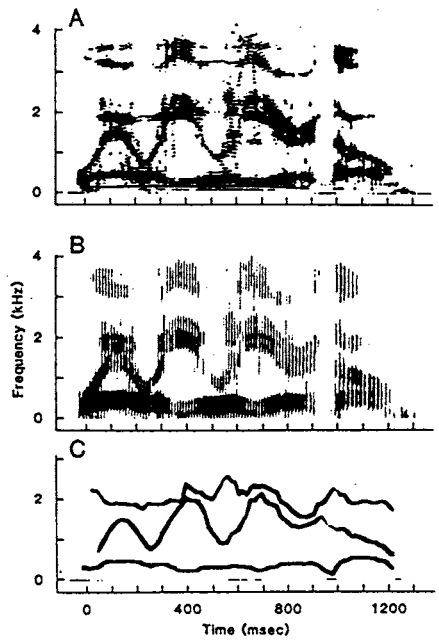

Fig. 1. (A) Narrowband spectrogram of the natural utterance "Where were you a year ago?" showing harmonic structure as narrow horizontal lines along the frequency scale. (B) Wideband spectrogram of the same utterance, showing formant pattern as dark bands along the time axis. The vertical striations correspond to individual laryngeal pulses. (C) Narrowband spectrogram of the three-tone sinusoidal replica. The energy concentrations follow the time-varying pattern of the formants above, but no energy is present except at the formant center frequencies. The amplitude variation in the sinusoidal pattern is not reproduced accurately.

0036-8075/81/0522-0947$00.50/0

ance. Overall, our three-tone pattern is a deliberately abstract representation of the time-varying spectral changes of the naturally produced utterance, although in local detail it is unlike natural speech signals.

The complex tone signal, having neither fundamental period nor formant structure, consists of none of those distinctive acoustic attributes that are traditionally assumed to underlie speech perception. None of the appropriate acoustic cues based on the acoustic events in natural speech is present in our stimuli. For example, there are no formant frequency transitions, which cue manner and place of articulation; no steady-state formants, which cue vowel color and consonant voicing; and no fundamental frequency changes, which cue voicing and stress (*6*). Similarly, the short-time spectral cues, which depend on precise amplitude and frequency characteristics across the harmonic spectrum, are absent from these tonal stimuli. An example would be the onset spectra that are often claimed to underlie perception of place features (*7*).

The perceptual importance of these attributes of speech signals has been indicated by theoretical models of sound production in the vocal tract. These models describe the speech signal as the product of a source and a filter (*3, 8*). Briefly, glottal pulsing provides a source in which energy is present at integral multiples of the fundamental frequency. The complex resonances of the pharyngeal, oral, and nasal cavities of the vocal tract are treated as a time-varying filter; the peaks in the vocal tract transfer function represent the formants. Perceptual tests of potentially distinctive attributes, however, have typically employed electronic or digital analogs of the source-filter theory of speech acoustics to create stimuli. In doing so, investigators have not questioned the necessity of harmonic spectra or broadband formant structure in speech perception, nor have they empirically raised the possibility that listeners attend to higher order relational properties of time-varying speech signals.

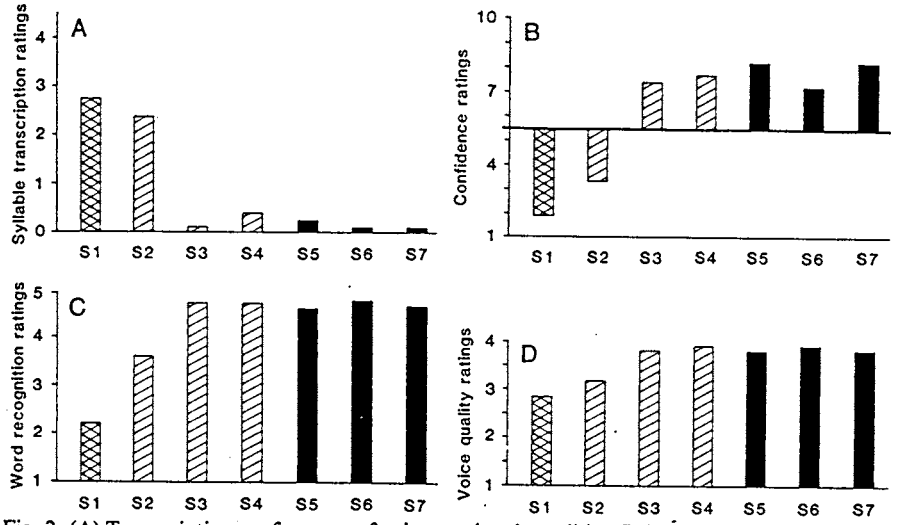

Fig. 2. (A) Transcription performance for instructional condition B (mean number of syllables correctly transcribed). (B) Detection ratings for instructional condition C (1 = confident yes, 10 = confident no). (C) Ratings of number of intelligible words in the tones (1 = every word, 2 = most, 3 = a few, 4 = almost none, 5 = none). (D) Voice quality ratings (1 = natural, 2 = peculiar, 3 = unnatural, 4 = nonspeech).

Table 1. Response categories and frequencies for group A listeners. Numbers in parentheses give the number of listeners making that type of response.

Stimulus condition	Response categories
S1 (T1 + T2 + T3)	Science fiction sounds (8), computer bleeps (5), music (4), several simultaneous sounds (3), human speech (3), "Where were you a year ago?" (2), radio interference (2), human vocalizations (1), artificial speech (1), bird sounds (1), reversed speech (1)
S2 (T1 + T2)	Science fiction sounds (7), computer bleeps (3), sirens (2), music (2), radio interference (2), tape recorder problems (1), reversed speech (1), whistles (1), artificial speech (1), human speech (1)
S3 (T2 + T3)	Science fiction sounds (14), radio interference (3), music (2), computer bleeps (2), whistles (1), several simultaneous sounds (1)
S4 (T1 + T3)	Science fiction sounds (9), artificial speech (5), computer bleeps (4), several simultaneous sounds (4), whistles (3), radio interference (2), tape recorder problems (2), human speech (1), human vocalizations (1), reversed speech (1), music (1)
S5 (T1)	Science fiction sounds (5), music (4), reversed speech (4), tape recorder problems (3), human speech (2), artificial speech (2), animal cries (2), bird sounds (2), radio interference (2), several simultaneous sounds (2), human vocalizations (1)
S6 (T2)	Sirens (7), bird sounds (4), mechanical sound effects (4), radio interference (4), animal cries (3), whistles (2), computer bleeps (1)
S7 (T3)	Bird sounds (17), whistles (6), mechanical sound effects (5), human vocalizations (3), human speech (1), artificial speech (1), computer bleeps (1), animal cries (1), music (1), radio interference (1), tape recorder problems (1)

The present study is a test of these assumptions. The absence of traditional acoustic cues to phonetic identity suggests that our sinusoidal replica of the sentence should be perceived as three independently changing tones. However, if listeners are able to perceive the tones as speech, then we may conclude that traditional speech cues are themselves approximations of second-order signal properties to which listeners attend when they perceive speech.

Our perceptual test consisted of three conditions in which independent groups of listeners were informed to different degrees about the tonal stimuli that they would hear (*9*). Within each instructional condition, different groups of 18 listeners each were assigned to seven stimulus conditions: the three tones presented together (S1: T1 + T2 + T3); three pairwise tone combinations (S2: T1 + T2, S3: T2 + T3, S4: T1 + T3); and each tone played separately (S5: T1, S6: T2, S7: T3). The three instructional conditions crossed with the seven stimulus conditions made 21 experimental conditions in all. In each condition a given sinusoidal pattern was presented four times in succession, at approximately 85-dB sound pressure level, by audiotape playback over matched and calibrated headphones.

In instructional condition A, listeners were asked simply to report their spontaneous impressions of the stimuli, having been told nothing about the nature of the sounds. Multiple responses were permitted. The accumulated responses, organized by stimulus condition, are displayed in Table 1. Apparently, the presentation of tones following the formant

center frequencies is insufficient to elicit phonetic perception; modal responses in each stimulus condition indicate that the majority of listeners did not hear the sinusoids as speech. A small number of responses in several conditions favored human- or artificial-speech interpretations, though, and two listeners in the three-tone condition responded that they heard "Where were you a year ago?" This outcome might be anticipated only if there were stimulus support of some kind for perceiving the linguistic content of these patterns. Even as a response to a direct request to generate a sentence in English, the probability of producing this sentence exactly is exceedingly small (*10*).

In instructional condition B, listeners were informed that they would hear a sentence produced by a computer and were asked to transcribe the synthetic utterance as faithfully as possible. We scored the responses in each condition for correct number of syllables transcribed (Fig. 2A). A large number of subjects identified the sentence in stimulus conditions 1 and 2. Nine of the listeners across these two conditions transcribed the entire sentence correctly; ten others reported that they could hear no sentence at all. The remaining listeners transcribed various numbers of syllables correctly.

We conclude from the first two instructional conditions that naïve listeners may not automatically perceive sinusoidal replicas of natural speech as linguistic entities. When instructed to do so, however, they perform well, presumably because the linguistic information, although not carried by acoustic elements producible by a vocal tract, is preserved in the time-varying relational structure of the stimulus pattern (*11*).

In instructional condition C, listeners were asked to evaluate the speech quality of the tonal stimuli. They were told that they would be presented with the sentence "Where were you a year ago?" and were asked to make several judgments. They were to report whether the sentence was discernible by responding yes or no; they were also to provide a confidence rating for their judgments, using a dual five-point scale. These responses were converted to a ten-point scale (Fig. 2B). In five of the stimulus conditions, listeners were very confident that they did not hear the sentence in the tones. However, in stimulus conditions 1 and 2, listeners were very confident that they recognized the intended sentence; the average confidence ratings in these two conditions did not differ significantly despite the absence of tone 3 in stimulus condition 2 (Scheffé post hoc means test, $P > .1$).

In the second task, listeners rated the number of words that could be identified in the particular pattern presented. As shown in Fig. 2C, the listeners could not identify any of the words in the sentence in five of the stimulus conditions. But in the three-tone condition, they reported that almost every word was clear. The omission of tone 3 from the pattern in stimulus condition 2 led listeners to report that significantly fewer words were intelligible ($P < .025$), yet this condition remains significantly different from stimulus conditions 3 through 7 ($P < .001$).

In the third task, listeners rated the voice quality of the stimuli [natural, funny (peculiar), unnatural, or nonspeech] (Fig. 2D). The split between stimulus conditions 1 and 2 and the other conditions is still quite evident, as it was in instructional condition B; however, these two stimulus patterns were judged to have unnatural voice quality despite their clear intelligibility. In essence, listeners apprehended the linguistic significance of the tonal patterns despite the radically unnatural, nonspeech quality (*12, 13*). That is, they were able to perceive the linguistic content of the utterance in the absence of acoustic patterns of the kind generated by the human vocal tract.

The results of this study cannot be explained within the framework of existing theories of speech perception (*14*), for the tones contained none of the elemental acoustic cues typically held to underlie speech perception. Although the tones presented information about formant center frequency, this minimal structure is evidently not sufficient to elicit phonetic perception spontaneously, as we saw in the performance of the naïve listeners. In fact, no property of the three-tone stimulus obliges the listener to hear it phonetically—except that its time-varying pattern of frequency change corresponds abstractly to the potential acoustic products of vocalization (*15*). The linguistically primed listeners were capable, for the most part, of directing their attention to the phonetic properties of the sinusoidal signal, merely by virtue of the instruction to listen in the speech mode of perception. For these subjects, the tones provided sufficient stimulation to evoke phonetic perception, albeit a kind that also identified the vocal source as unnatural.

We conclude, then, that speech perception can endure the absence of particular short-time acoustic spectra and traditional formant-based acoustic cues only insofar as the pattern of change in the natural signal is preserved over transposition from harmonic to sine wave spectra (*16*). Further examples of nonspeech tonal analogs of natural speech are needed to characterize more precisely the time-varying relations that support phonetic perception.

ROBERT E. REMEZ
Department of Psychology, Barnard College, Columbia University, New York 10027

PHILIP E. RUBIN
Haskins Laboratories, New Haven, Connecticut 06510

DAVID B. PISONI
THOMAS D. CARRELL
Department of Psychology, Indiana University, Bloomington 47405

References and Notes

1. C. G. M. Fant, *Logos* 5, 3 (1962); A. M. Liberman, F. S. Cooper, D. P. Shankweiler, M. Studdert-Kennedy, *Psychol. Rev.* 74, 421 (1967); I. G. Mattingly, *Am. Sci.* 60, 327 (1972); K. N. Stevens and S. E. Blumstein, *J. Acoust. Soc. Am.* 64, 1358 (1978).
2. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech* (Springer, New York, 1976).
3. T. Chiba and M. Kajiyama, *The Vowel: Its Nature and Structure* (Tokyo-Kaiseikan, Tokyo, 1941).
4. C. G. M. Fant, *The Acoustic Theory of Speech Production* (Mouton, The Hague, 1960). The closely spaced horizontal lines shown in Fig. 1A are the harmonics of the fundamental frequency of phonation, and are typically revealed in narrowband spectrograms.
5. Typically, the amplitude of the valleys in the spectrum of natural speech ranges from 10 to 30 dB below the amplitude of the peaks (*7*).
6. A. M. Liberman and M. Studdert-Kennedy, in *Handbook of Sensory Physiology*, vol. 3, *Perception*, R. Held, H. Leibowitz, H.-L. Teuber, Eds. (Springer, New York, 1978).
7. K. N. Stevens and S. E. Blumstein, in *Perspectives in the Study of Speech*, P. D. Eimas and J. L. Miller, Eds. (Erlbaum, Hillsdale, N.J., in press).
8. K. N. Stevens, in *Handbook of Physiology-Respiration I*, W. O. Fenn and H. Rahn, Eds. (American Physiological Society, Washington, D.C., 1964).
9. Our listeners were students of introductory psychology at Indiana University, Bloomington. They were unfamiliar with synthetic speech.
10. G. A. Miller and N. Chomsky, in *Handbook of Mathematical Psychology*, R. D. Luce, R. R. Bush, E. Galanter, Eds. (Holt, New York, 1960), vol. 2.
11. It has often been emphasized that a variety of acoustic events may cue a single phonetic feature in the absence of other, redundant cues; experiments with synthetic speech in which phonetic distinctions were minimally cued indicate that listeners tolerate schematized speech signals with little loss of understanding [A. M. Liberman and F. S. Cooper, in *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, A. Valdman, Ed. (Mouton, The Hague, 1972)]. For this reason, listeners probably do not require stimuli to display the acoustic "stigmata" of speech to be candidates for phonetic interpretation [A. M. Liberman, I. G. Mattingly, M. T. Turvey, in *Coding Processes in Human Memory*, A. W. Melton and E. Martin, Eds. (Winston, Washington, D.C., 1972)]. However, even schematized synthetic speech has consisted of acoustic cues that are utterable in principle as components of a speech signal; these cues have specific articulatory rationales. This resemblance of schematized synthetic speech to natural speech may have led theorists to underestimate the abstractness of the stimulus properties relevant to perception. Signals consisting of sinusoids may be used to study these more abstract time-varying acoustic properties underlying phonetic perception, for their phonetic effects cannot be explained by arguing that they are components of natural speech signals, or by arguing that they are acoustic products of natural vocal articulation.

12. Although much intelligible synthetic speech would also be judged unnatural, this may be ascribed to the practice of presenting the speech cues in contexts of minimal variation in the acoustic parameters that are irrelevant to intelligibility, but which affect speech quality nonetheless [Liberman and Cooper, in (*11*)]. A synthesizer that produces a harmonic spectrum, broadband formants, and a fundamental period within the normal range will sound unnatural, and perhaps be unintelligible, despite the acoustic resemblance to natural speech if the synthesis of prosodic variation—of speech rhythm, meter, and melody—is inappropriate [J. Allen, *Proc. IEEE* 64, 433 (1976)]. The judgment that this kind of synthetic imitation of speech signals is unnatural is, therefore, quite different from the judgment of unnaturalness in the present case.
13. Although the intelligibility of our sinusoidal sentence is predicted by the co-occurrence of tones 1 and 2 but not of tones 1 and 3, the effectiveness of each tone pair will vary as a function of the phonetic composition of the particular utterance. While the resonance associated with the oral cavity is thought to be primary for phonetic perception [G. M. Kuhn, *J. Acoust. Soc. Am.* 65, 774 (1979)], either the second or third formant may be affiliated with the oral cavity, depending on the phone in question [K. N. Stevens, in *Human Communication: A Unified View*, E. E. David and P. B. Denes, Eds. (McGraw-Hill, New York, 1972)]. Therefore, the critical tone pair will sometimes include tone 2 and sometimes tone 3, depending on the phonetic composition of the utterance.
14. The proposal that listeners "track" formant frequency variations can be entertained as an explanation of our findings only if the meaning of the term formant is extended to mean "any peak in the spectrum." In its present sense, formant refers to a natural resonance of the vocal cavities [L. Hermann, *Arch. Gesamte Physiol.* 58, 264 (1894)]. Quite literally, then, there are no vocal resonances in our tone complexes (although listeners who succeed in extracting the meaning probably do so because the tones preserve time-varying properties of vocally produced signals). Our preference is to retain the literal meaning of formant and to conclude, therefore, that the difference between voiced speech signals and the tonal signals is that the former contain broadband formant structure and harmonic spectra while the latter contain merely inharmonic peaks with infinitely narrow bandwidths.
15. Our finding is related, in some sense, to early studies of "vowel pitch," in which simple steady-state tones were judged to possess "vocality," or speechlike qualities [W. Kohler, *Z. Psychol.* 58, 59 (1910); J. D. Modell and G. J. Rich, *Am. J. Psychol.* 26, 453 (1915); E. B. Titchener, described in E. G. Boring, *Sensation and Perception in the History of Experimental Psychology* (Holt, New York, 1942), p. 374]. More recent studies have shown that listeners may identify brief complex sinusoidal patterns as isolated syllables, and therefore as speech sounds, when they are supplied with restricted response alternatives in judgment tasks with low uncertainty [J. E. Cutting, *Percept. Psychophys.* 16, 601 (1974); P. J. Bailey, A. Q. Summerfield, M. Dorman, *Haskins Laboratories Status Report on Speech Research, SR-51, 52* (1977), p. 1; C. T. Best, B. Morrongiello, R. Robson, *Percept. Psychophys.*, in press; M. E. Grunke and D. B. Pisoni, in *Proceedings of the Ninth International Congress of Phonetic Sciences* (Institute of Phonetics, Copenhagen, 1979), vol. 2, p. 461]. The present study, however, makes use of neither a closed response set nor a judgment task with low uncertainty to obtain the effect of intelligibility.
16. We recently synthesized "A yellow lion roared," thereby extending the range of tone synthesis to nasal manner as well as the stop consonant, liquid consonant, and vowel phone classes represented here. Similar findings have been obtained with this sentence, indicating that the present results are not due to peculiarities of the sentence used in these tests.
17. We thank C. Marshall, J. Montelongo, and S. Gans for their assistance. We also thank D. Aslin, P. Bailey, A. Liberman, and F. Restle, among many others, for very helpful advice and comments on an earlier version of the manuscript. This research was supported by grants MH 32848 (R.E.R.) and MH 24027 (D.B.P.) from the National Institute of Mental Health and by grant HD-01994 from the National Institute of Child Health and Human Development to Haskins Laboratories.

4 August 1980; revised 23 December 1980