Perception & Psychophysics
1990, 48 (4), 313-325

On the perception of speech from time-varying acoustic information: Contributions of amplitude variation

723

ROBERT E. REMEZ
Barnard College, New York, New York

and

PHILIP E. RUBIN
Haskins Laboratories, New Haven, Connecticut

The cyclic variation in the energy envelope of the speech signal results from the production of speech in syllables. This acoustic property is often identified as a source of information in the perception of syllable attributes, though spectral variation can also provide this information reliably. In the present study of the relative contributions of the energy and spectral envelopes in speech perception, we employed sinusoidal replicas of utterances, which permitted us to examine the roles of these acoustic properties in establishing or maintaining time-varying perceptual coherence. Three experiments were carried out to assess the independent perceptual effects of variation in sinusoidal amplitude and frequency, using sentence-length signals. In Experiment 1, we found that the fine grain of amplitude variation was not necessary for the perception of segmental and suprasegmental linguistic attributes; in Experiment 2, we found that amplitude variation was nonetheless effective in influencing syllable perception, and that in some circumstances it was crucial to segmental perception; in Experiment 3, we observed that coarse-grain amplitude variation, above all, proved to be extremely important in phonetic perception. We conclude that in perceiving sinusoidal replicas, the perceiver derives much from following the coherent pattern of frequency variation and gross signal energy, but probably derives rather little from tracking the precise details of the energy envelope. These findings encourage the view that the perceiver uses time-varying acoustic properties selectively in understanding speech.

In discussions of speech perception, the stability of the listener's phonetic impressions has often been contrasted with the variability of the underlying acoustic patterns (Fant, 1962; Fowler & Smith, 1986; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). These descriptive and theoretical accounts have been aimed at establishing the principles that govern the acoustic realization of phonetic segments intended by the talker, and the principles that in turn govern the perceptual realization of phonetic impressions by the listener. Though knowledge of the speech signal has developed greatly through this search for acoustic cues to phonetic identity, a satisfactory account of the listener's perceptual finesse eludes us. The difficulty may be traced to two critical facts about speech: There does not seem to be a core set of acoustic cues (Liberman & Cooper, 1972); and the variability of the signal structure does not appear to indicate a normal set of acoustic elements about which variation occurs (Bailey & Summerfield, 1980).

Although accounts of perception based on discrete acoustic cues and cue combinations retain a straightforward and durable appeal (Massaro, 1987; Stevens & Blumstein, 1981), some investigators have taken variability to be central in describing speech perception by the human listener, and they have therefore pursued the hypothesis that speech perception, however it may be embodied, depends in part on coherent time-varying properties of the signal (e.g., Best, Studdert-Kennedy, Manuel, & Rubin-Spitz, 1989; Jenkins, Strange, & Edman, 1983; Kewley-Port & Luce, 1984). Our studies have likewise supported the claim that the listener can draw phonetic information from speech signal variation, in some sense independently of the specific acoustic elements composing the signal, and independently of specific auditory impressions of speech sounds (Remez & Rubin, 1983; Remez, Rubin, Pisoni, & Carrell, 1981). In this research, we have employed a sinusoidal replication technique, in which synthetic acoustic patterns are matched to gross spectral changes, but not to fine acoustic details, of a speech signal. Perceptual tests with these materials provide evidence of the role of time-varying information in speech perception.

The authors gladly acknowledge the generous advice and stern encouragement of Jessica Lang, Paula Payton, Jan Rabinowitz, Amanda Steinberg, Ress Young, and the anonymous reviewers of an earlier submission of Experiment 1. This research was supported by grants from NIDCD (00308) to R. E. Remez, and from NICHHD (01994) to Haskins Laboratories. Please address correspondence to R. E. Remez, Department of Psychology, Barnard College, 3009 Broadway, New York, NY 10027-6598. Internet: remez@cunixb.cc.columbia.edu.

In our studies of perceptual susceptibility to phonetic information carried by sinusoidal replicas of speech signals, three or four pure tones have been used to represent the frequency and amplitude variation of oral, nasal, and fricative formants of natural utterances. The result is a signal that differs drastically in its physical properties from natural speech. Typically, speech exhibits some consistent acoustic attributes: a pulse train and a harmonic series arising from the glottal excitation; broadband natural resonances of the supralaryngeal vocal tract, the formants; and aperiodic elements: release bursts, low-energy aspirate formants, and turbulence in fricative formants—to note three. Sinusoidal replicas of speech lack this fine grain of acoustic detail on which most characterizations of perceptual cues implicitly rely, yet they are comprehensible despite the absence of these familiar acoustic products of vocalization (Remez et al., 1981). (See Figure 1.) On the basis of studies employing sinusoidal replicas of natural utterances, and in agreement with related results found with more conventional acoustic techniques, we have learned that the perception of utterances may rely as much on the patterns composed by the elementary acoustic cues as on the specific psychoacoustic effects of the signal elements themselves.

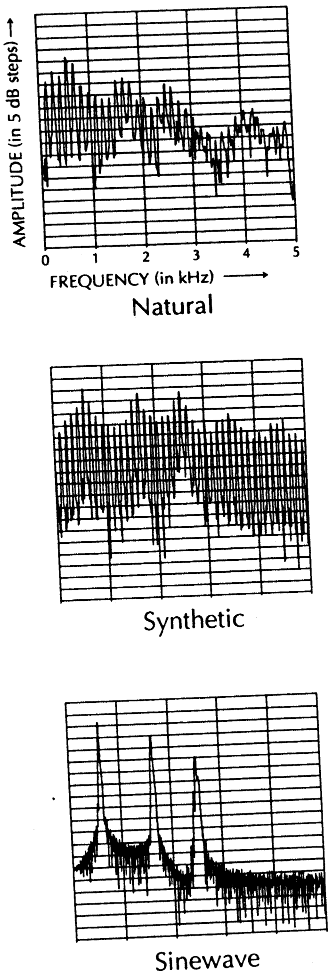


Figure 1. The three panels present short-time spectra of three kinds of signal. Top: natural speech. Middle: cascade-type synthetic speech. Bottom: sinusoidal replication of speech.

Although our tests have suggested that the frequency variation of a sinusoidal sentence is perceptually effective, this interpretation can be challenged. Because each tonal component has represented both the frequency and the amplitude variation of a natural resonance, a signal of this description delivers veridical formant amplitude information on anomalous carriers, which are several simultaneously varying sinusoids. In contrast to the predictions of our original hypothesis, which emphasized the coherent variation in tone frequency, we may alternatively find that tone amplitude variation plays a prominent perceptual role when sinusoidal signals are heard phonetically. In fact, a number of views of the perception of fluent natural speech propose a processing stage in which lexical patterns are hypothesized from metrical structure, and especially so when phonetic information is defective or ambiguous (see, e.g., Cutler & Foss, 1977; Cutler & Norris, 1988; Huggins, 1978; Nakatani & Schaffer, 1978). Perhaps a subject who attempts to transcribe a sinusoidal sentence employs such guesswork after exploiting the amplitude variation to perceive the sentence meter and rhythm.

If this is true of the perception of sinusoidal sentences, then the listener does not perceive the message as we have claimed, from phonetic information preserved primarily in the time-varying pattern of tone frequencies. Moreover, this alternative to our frequency-based claim seems especially appealing, considering that the sensory derivation of this aspect of the signal, the envelope of a sinusoidal replica, probably occurs in the same way as it does for a natural signal, by summing the energy across spectral components. This encourages the possibility that the veridical amplitude envelope of a sinusoidal signal supplies phonetic information.

Accordingly, in the three experiments reported here, we attempted to resolve this issue by determining the perceptual effectiveness of tone amplitude variation in providing phonetic information within the context of sinusoidal replication. Experiment 1 demonstrated that the perceiver can do without this acoustically veridical property of sinusoidal sentence replicas under some circumstances, as our initial frequency-based hypothesis claimed; Experiment 2 solved part of the puzzle by establishing the sufficiency of tone frequency and amplitude variation as information in the perception of syllabic if not segmental properties; and Experiment 3 revealed that gross changes in the energy envelope, as opposed to the fine structure of the envelope, can have a great influence on the perception of speech from time-varying information.

EXPERIMENT 1

In the first study, we sought to determine the effect of veridical amplitude variation in the perception of sinusoidal sentences. The test called for two kinds of perceptual report from subjects, under four conditions of sinusoidal replication designed to identify the sufficiency of variation in sinusoidal frequency and amplitude.

The subjects were asked both to transcribe a sentence and to report the number of syllables that occurred. In the first condition, a sinusoidal pattern presented both the resonance frequency and the amplitude of each of the lowest three formants of a natural utterance. In the second condition, listeners heard a pattern that preserved the frequency variation of the first three formant centers, but it was presented at a constant level of energy throughout its duration. In the third condition, the sinusoidal pattern preserved the frequencies of the first three formant tracks, but with a misleading amplitude contour imposed on it. In the fourth condition, the sinusoidal pattern had the natural formant amplitude variation, but the tones in this case exhibited constant frequency, so that the effect of the natural amplitude contour could be estimated without concurrent frequency variation.

If the amplitude variations of the sinusoidal signal provided information about the prosodic structure of the utterance, and if the subjects relied primarily on this source of information in performing the two tasks that we set for them, then syllable counting would be accurate in the first and the fourth conditions, and poorer in the second and third conditions. If subjects perceived the phonetic sequence on the basis of the time-varying properties of frequency variation, however, transcription and syllable counting would be accurate in all conditions but the last, in which there was no frequency variation.

Method

Subjects. Fourteen adults with normal hearing in both ears made up each of the four groups that were tested, blocked by the four synthesis conditions. The subjects were drawn from sections of introductory psychology classes, and they received course credit for their participation. All were native speakers of English, and none had participated in any other experiments in which sinusoidal signals had been employed.

Acoustic test materials. Four different sinusoidal patterns produced by the sine wave synthesizer at Haskins Laboratories were used. This software synthesizer accepts frequency, amplitude, and duration values for setting the parameters of digital oscillators, and it calculates the resulting waveform with 12-bit amplitude resolution at a designated frame rate. A sentence ("Where were you a year ago?") uttered by one of the authors was used as the model for the four sinusoidal patterns.

The natural utterance was recorded on audiotape (Scotch No. 208) with a high-quality voice microphone (Shure SM87) in a small sound-attenuating chamber (IAC Model 400A), and then converted to digital records by filtering it (4.5-kHz low-pass, −40-dB/octave rolloff) and sampling it at 10 kHz, using a PCM system implemented on a DEC VAX-11/780. The sampled natural speech data were then analyzed with the technique of linear prediction, to determine the center frequency and the amplitude of each of the lowest three formants. Formant frequencies and amplitudes were estimated at 10-msec intervals throughout the utterance, and these derived values were then appropriated for use as a table of sinusoidal synthesis specifications.

The four patterns that we constructed are shown in Figure 2, and they correspond to the four perceptual tests of sinusoidal replication. Figure 2A shows the pattern used in the first test, which contained three sinusoids following the frequency and amplitude variations of the original natural utterance. Figure 2B shows the pattern tested in the second condition, which preserved the frequency variation of the spectrum, but with each tone at constant power throughout the duration of the sentence. The second tone had approximately half the power of the first, and the third had approximately half that of the second. Figure 2C shows the signal that was used in the third test, in which a misleading artificial amplitude pattern was imposed on each tone, though the frequency variations of each tone follow the values derived from the natural model. Last, Figure 2D shows the pattern of the fourth test, exhibiting natural amplitude values carried by constant-frequency tones, the lowest at 500 Hz, the second at 1500 Hz, and the third at 2500 Hz.

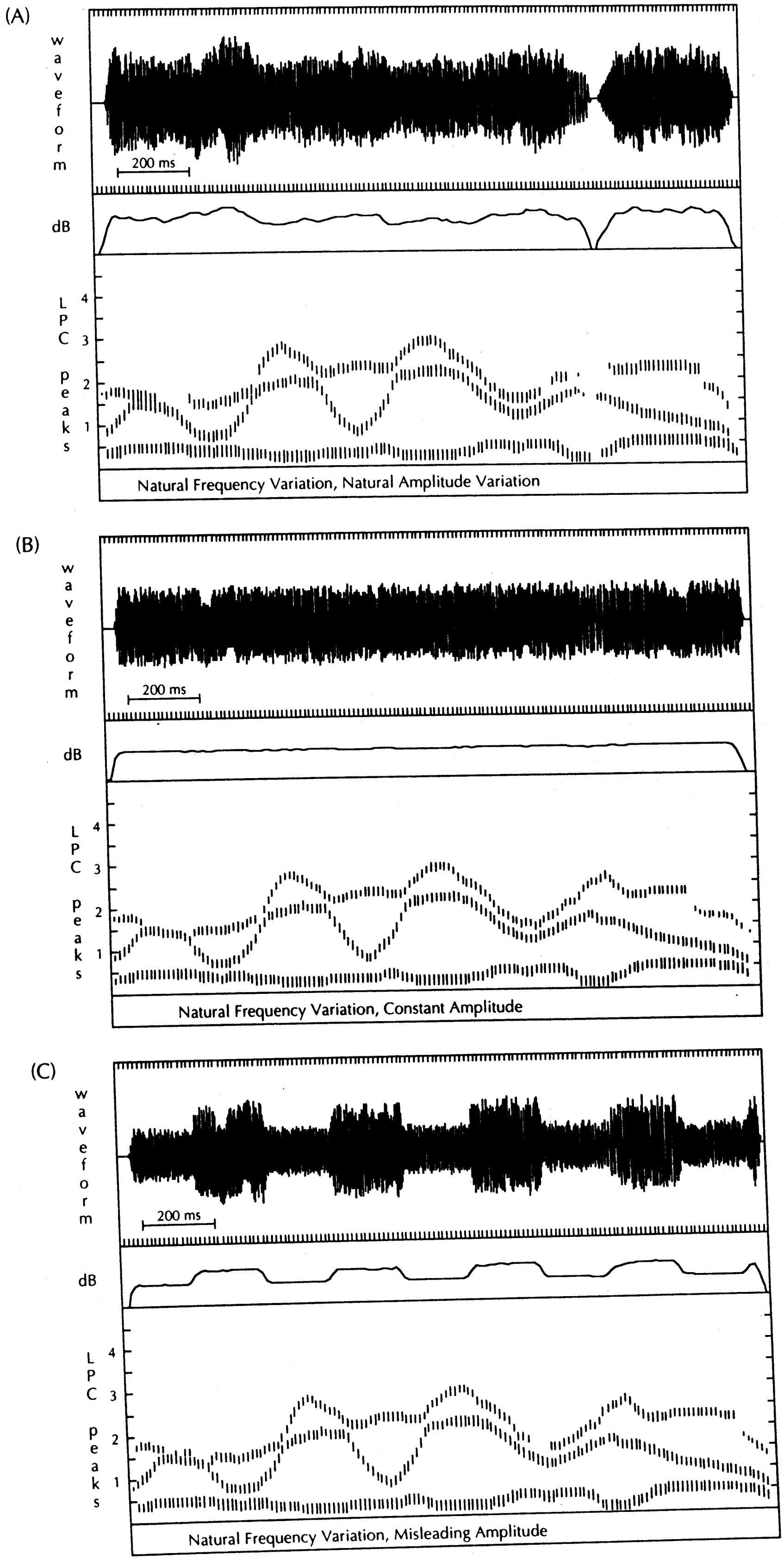
The synthetic sinusoidal patterns were converted from digital records to analog signals, recorded on half-track .25-in. audiotape (Scotch No. 208) at Haskins Laboratories, and were presented to subjects via tape playback using an Otari MX-5050B tape recorder and a Crown D74 power amplifier. Listeners sat in carrels in a sound-shielded room, and signals were presented binaurally at an approximate level of 65 dB SPL over matched and calibrated Telephonics TDH-39 headsets.

Procedure. This experiment was composed of four tests, each with different listeners. Each test corresponded to one of the four sinusoidal signals shown in Figure 2: (1) natural frequency and amplitude variation; (2) natural frequency variation, with constant amplitude; (3) natural frequency variation, with artificial (and presumably) misleading amplitude variation; and (4) constant-frequency sinusoids with natural amplitude variation. A single sentence was presented four times in succession, separated by 1 sec, at the conclusion of which the subjects reported by writing their responses in prepared booklets.

Listeners were tested in groups of 6 or fewer. They were briefly instructed that synthetic speech was to be presented over the headphones, and they were asked to mark an answer sheet with (1) their impression of the number of syllables in the computer's utterance, and (2) a transcription of the sentence that the computer produced.

Results

The outcome of the tests, shown in Figure 3, was straightforward. The transcription test revealed that subjects were not hindered by defective amplitude properties of the signal as long as the information carried by tone frequency variation was available. Transcription performance for three tests was scored as the number of syllables correctly reported, with a maximum of 7 if the sentence was transcribed completely. The group performance levels for the tests of the effect of amplitude envelope when tone frequency varied concurrently are shown in the top panel of Figure 3. (No transcriptions were scored for the fourth test, in which tone frequency did not vary, and in which case the subjects did not report hearing phonetic segments.) These performance levels exceed a score of 0 syllables transcribed correctly, and they do not differ significantly from each other, as was shown in two statistical tests. First, the analysis of variance performed on these data revealed no significant effect of amplitude envelope on the accuracy of transcription [$F(2,36) = 0.64, p > .5$]. Second, the grand mean differed significantly from a score of zero [$t(38) = 11.19, p < .001$]. Because some subjects failed to follow the test instructions, we were unable to include the reports of 1 in the second test, and of 2 in the third.



(A) Natural Frequency Variation, Natural Amplitude Variation

(B) Natural Frequency Variation, Constant Amplitude

(C) Natural Frequency Variation, Misleading Amplitude

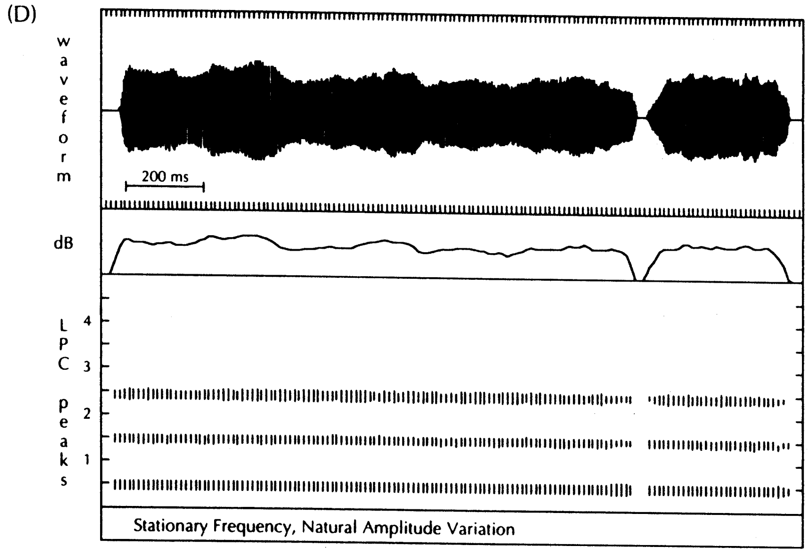


Figure 2. (A, opposite page) The waveform (top), rms energy in decibels (middle), and spectral analysis for the natural amplitude test item (bottom). (B, opposite page) The waveform (top), rms energy in decibels (middle), and spectral analysis for the constant amplitude test item (bottom). (C, opposite page) The waveform (top), rms energy in decibels (middle), and spectral analysis for the misleading amplitude test item (bottom). (D) The waveform (top), rms energy in decibels (middle), and spectral analysis for the test item possessing natural amplitude variation imposed on stationary-frequency tones (bottom).

The results of the syllable counting test were surprising; they indicated that the perception of syllabic attributes of the sinusoidal utterance depended less on the amplitude envelope than on frequency variation. Group averages for reports of syllable counts in the four conditions are shown in the bottom panel of Figure 3. The analysis of variance performed on these data showed a significant effect of the acoustic conditions [$F(3,48) = 42.91$, $p < .001$], and a Scheffe post hoc means test confirmed the difference ($p < .001$) between the fourth test (stationary frequency, concurrent natural amplitude variation) and the other three tests (natural frequency and amplitude variation, natural frequency variation at constant power, and natural frequency variation with misleading amplitude variation).

Discussion

On the basis of this first experiment, it seems that a sinusoidal replica preserves useful acoustic information for the phonetic composition of an utterance solely in its tone frequency variation. We may reject the hypothesis that the listener is only able to transcribe the phonetic properties of sinusoidal sentences by combining prosodic information carried in the veridical variation of the sinusoidal amplitude envelope with ambiguous or incomplete information supplied by the anomalous sinusoidal carriers. The evidence on this point provided by our listeners is compelling. They were unable to follow the syllable structure of the utterance in the crucial fourth condition, when the natural amplitude variation was presented without concurrent frequency variation. In that condition, the single salient linguistic property conveyed by the energy envelope was apparently the closure of the vocal tract during the production of a stop consonant, which may explain the consistent report that the pattern consisted of two syllables—one preclosure, and one postclosure.

A minimal role of energy tracking in speech perception is also consistent with the results of the third condition, in which subjects were able to apprehend the phonetic detail even when the energy contour was grossly inappropriate to the segments subordinate to it. It seems that listeners who transcribed these sinusoidal replicas of speech must have relied on information about the phonetic sequence available primarily in the frequency variation, as we have claimed from the outcome of our prior research (Remez, 1987; Remez & Rubin, 1983, 1984; Remez, Rubin, Nygaard, & Howell, 1987; Remez et al., 1981).

Although this study revealed that listeners may disregard a rather consistent acoustic correlate of the syllable pattern, it does not permit us to draw a strong and general conclusion about phonetic perception from time-varying sinusoidal replicas, nor about ordinary speech perception. Though we have shown that veridical amplitude variation is not essential in perceiving tonal analogues of speech, we have not shown that amplitude variation is disregarded when it is available. Perhaps there are differences across utterances in the effectiveness of amplitude information, and in Experiment 1 we may have employed an utterance in which the correspondences of energy envelope and syllable structure were less than straightforward. In any case, a fair test of amplitude as a source of information requires diversity in the linguistic properties of

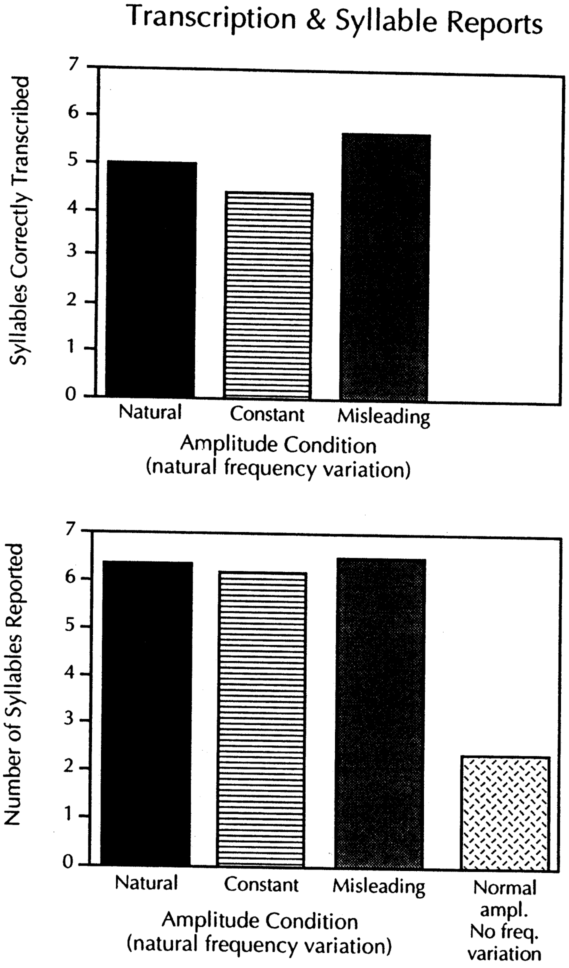


Figure 3. Group performance on the transcription task (top) and on the syllable counting task (bottom) of Experiment 1.

the test materials. The phone classes comprised by the sentence used in Experiment 1 were convenient but not representative of the inventory of English, consisting of liquid consonants, vowels, and a single stop consonant. Ordinary speech is not similarly restricted, and a general claim requires a less restricted test.

In Experiment 2 we therefore employed a sentence with nasal consonants, voiceless stops, voiced and voiceless fricatives, and consonant clusters, to remove this limitation on the first experiment. These segments are distinguished in the amplitude envelope to different degrees; stops and affricates especially are usually marked in the waveform with silence (Halle, Hughes, & Radley, 1957), and they may be viewed as the likeliest kind of phonetic segment to depend perceptually on information provided by amplitude. By adopting a factorial design, we also expected to clarify the relative independent contributions of frequency and amplitude variation to intelligibility and to judgments of syllable numerosity.

EXPERIMENT 2

The objective of Experiment 2 was to provide a more general test of the findings of the first experiment. Although it appeared that perceivers were indifferent to the only acoustically veridical component of a sinusoidal replica, our test had involved a sentence and a set of conditions that might have obscured the independent perceptual roles of tone amplitude variation and tone frequency variation. To provide the remedy, two aspects of the design were modified in Experiment 2. First, a different sentence was used, in order to replicate the acoustic treatments of Experiment 1 with greater phonetic and acoustic variety. Second, an extended set of conditions was used to examine cases of natural, constant, and misleading energy envelopes imposed on a set of natural, constant, and misleading frequency patterns, in a factorial design. This combination produced a test that was less phonetically restricted, in which the potential contributions of tone frequency and energy could be assessed. Again, we appraised the impact of these acoustic properties on perception of the phonetic and syllable pattern by collecting two reports on each trial: (1) a report of the number of syllables in the utterance, and (2) a transcription of the words.

Method

Subjects. Two hundred and twenty-two subjects were tested overall, in nine test conditions. Some listeners were paid for participating, while others were drawn from the introductory psychology subject pool. Each reported normal hearing in both ears, and none had participated in any other experiments in which sinusoidal signals had been employed.

Acoustic test materials. Nine sinusoidal sentence patterns were created on the model of a natural utterance ("My t.v. has a twelve-inch screen") produced by one of the authors. As in the case of Experiment 1, this utterance was recorded on tape, low-pass filtered at 4.5 kHz, and sampled by the VAX at 10 kHz with 12-bit resolution. The digital records of the utterance were analyzed at 10-msec intervals with the technique of linear prediction, to derive frequency and amplitude estimates of the oral, nasal, and fricative formants. After several erroneously estimated frequency values had been corrected, these analysis data were used to compose a table of sinusoidal synthesis specifications.[1] Incorporating both natural frequency and amplitude variation, this synthesis parameter set was used to make the *natural frequency, natural amplitude* test item, which is illustrated in Figure 4A.

Eight variant patterns were then composed by modifying the synthesis parameters that had preserved the natural frequency and amplitude values. In all, three different kinds of frequency variation—the natural values, a misleading frequency pattern, and a stationary frequency pattern—were crossed with three kinds of amplitude variation—the natural values, a misleading amplitude pattern similar to that used in Experiment 1, and an unchanging amplitude pattern. The construction of each of the variants was straightforward:

Natural frequency, constant amplitude. To impose constant amplitude while preserving natural frequency variation, the amplitudes of Tones 1, 2, and 3 were each fixed throughout the utterance. A natural rolloff of −6 dB/octave was approximated across the three tones, replicating the oral resonances. This manipulation also maintained the constant level through the portions of the sinusoidal replica that corresponded to stop closures. In addition, the intermittent and

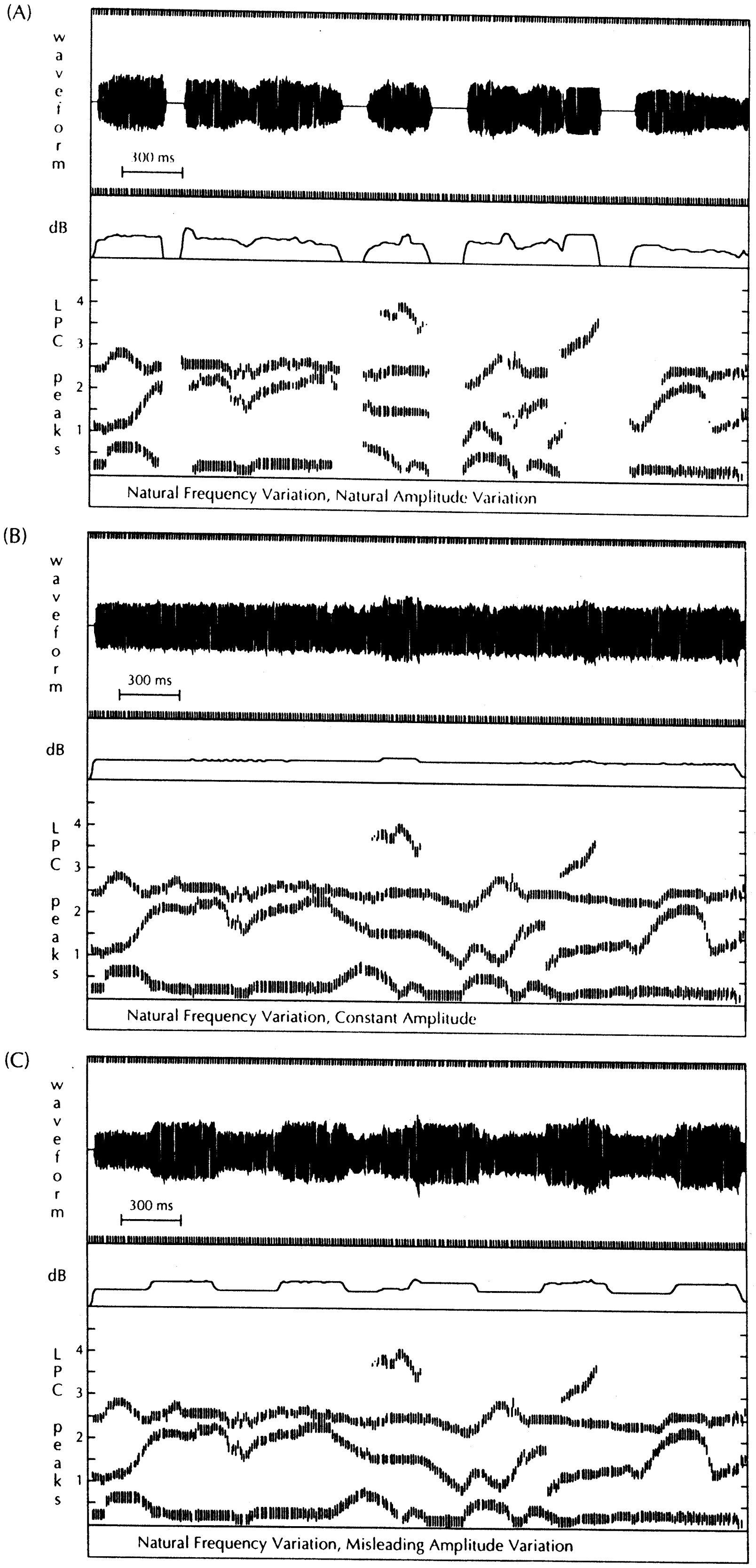


Figure 4. (A) Natural frequency, natural amplitude test item, sentence "My t.v. has a twelve-inch screen." (B) Natural frequency, constant amplitude item. (C) Natural frequency, misleading amplitude test item.

brief introduction of a fourth tone, correlated with the /z/ frication in [hæz] and the affricate-fricative /čs/ sequence in [ɪnčskrin] was presented similarly by imposing a constant energy level during the duration of each brief tone. See Figure 4B for an illustration of this item.

Natural frequency, misleading amplitude. To impose an arbitrarily misleading energy pattern while preserving natural frequency variation, the synthetic pattern was divided conceptually into 310-msec portions, every other one of which was created with a 20% increase in tone amplitude relative to the instance of constant frequency. This produced 10 individual amplitude episodes, or five paired sequences. Again, the tones maintained the energy through portions of the utterance corresponding to stop closures. See Figure 4C for an illustration of this item.

Misleading frequency, natural amplitude. To compose a pattern of misleading tone frequency values, we imposed a sinusoidal modulation on each of four component tones, centered respectively at 500, 1500, 2500, and 3500 Hz, with a maximum excursion of ±250 Hz, and a period of modulation of 500 msec. This resulted in a four-tone frequency pattern that cycled approximately two and a half times over the duration of the signal. Again, this pattern of frequency values was combined with the formant amplitude values derived from the natural utterance, resulting in a sinusoidal pattern exhibiting misleading frequency variation and natural amplitude variation.

Misleading frequency, constant amplitude. The pattern of sinusoidally modulated tone frequencies was combined with constant-level amplitude values. The result was the pattern shown in Figure 5A, which also exhibited a natural rolloff.

Misleading frequency, misleading amplitude. Amplitude portions differing in energy by 20% again alternated throughout the duration of this signal. This misleading amplitude pattern was combined with the sinusoidally modulated tone frequency values.

Constant frequency, natural amplitude. To remove natural frequency variation while preserving natural amplitude variation, the tone frequency values specified in the natural frequency, natural

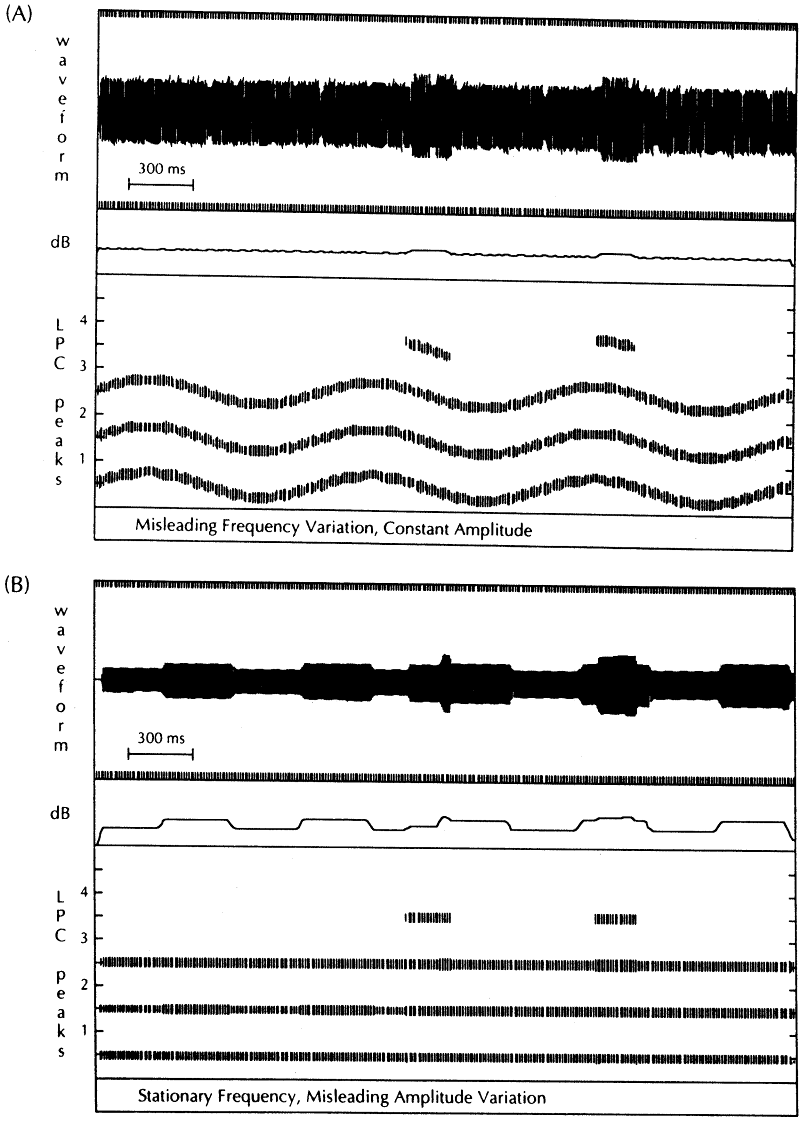


Figure 5. (A) Misleading frequency, constant amplitude test item. (B) Stationary frequency, misleading amplitude test item.

amplitude item were set to 500, 1500, and 2500 Hz, in the respective cases of Tones 1, 2, and 3, and to 3500 Hz for the brief intermittent "fricative" Tone 4.

Constant frequency, constant amplitude. The synthesis values eliminated all frequency and amplitude variation by combining stationary frequency values of 500, 1500, 2500, and 3500 Hz for the four component tones and constant amplitude levels. Again, the natural rolloff of −6 dB/octave was imposed across the tones.

Constant frequency, misleading amplitude. The amplitude pattern again contained alternating portions differing in energy by 20%, imposed on tone frequency values that were stationary throughout the duration of the signal. Figure 5B shows this test item.

The digital records of the synthetic sinusoidal patterns were converted to analog signals and recorded on audiotape. Listening tests were presented via tape playback. The subjects sat in carrels in a sound-shielded room, wearing matched and calibrated Telephonics TDH-39 headsets, and heard the signals binaurally attenuated to a level of approximately 65 dB SPL.

Procedure. There were nine conditions in this experiment, corresponding to the nine sinusoidal patterns varying in frequency and amplitude properties. The subjects were assigned randomly to conditions, and were tested in groups of 6 or fewer. As was the case in Experiment 1, subjects were instructed that synthetic speech was to be presented over the headphones, and were asked to report both an impression of the number of syllables in the computer's utterance, and also to transcribe the sentence that the computer produced.

Each session began with an eight-sentence pretest that was used both to promote the subject's susceptibility to sinusoidally replicated utterances and, later, to assess that susceptibility. The sentences were drawn from the set of Egan (1948). At the immediate conclusion of the pretest, one of the nine test sentences was presented. Every sentence was presented eight times in succession, separated by 3 sec, with 10 sec between trials. The subjects reported their impressions by writing in prepared booklets.

Results

Pretest. The results of the eight-sentence pretest were used to determine whether the subjects were susceptible to the phonetic attributes of tone analogues of speech. In the present case, we found that 18% of the subjects (42 of them) failed to transcribe any sentences on the pretest, and they were therefore eliminated from the data set. This left 20 subjects in each of nine test conditions who passed the pretest and who presumably were sensitive to the phonetic effects of sinusoidal replication of speech.

Transcription performance. The statistical analysis of transcription reports was unnecessary, because there was only one condition in the set of nine in which transcription occurred: the combination of natural frequency and natural amplitude. With that congruence of acoustic properties, the average number of syllables reported correctly was 4.7. In neither of the two other conditions in which natural frequency values were presented, nor in the six conditions in which misleading or unchanging frequency values were presented, did average transcription performance exceed 1 syllable.

Syllable numerosity reports. The results of the tests of acoustic influence on judgments of syllable numerosity were determined by a two-way analysis of variance, crossing three frequency categories with three amplitude categories. The group means of apparent syllable reports are shown in Figure 6, in each of the nine conditions. Both main effects were observed: for frequency [$F(2,171) = 15.88, p < .001$] and for amplitude [$F(2,171) = 8.50, p < .001$]. Moreover, the interaction of these factors was also significant [$F(4,171) = 7.86, p < .001$], as is apparent from the figure.

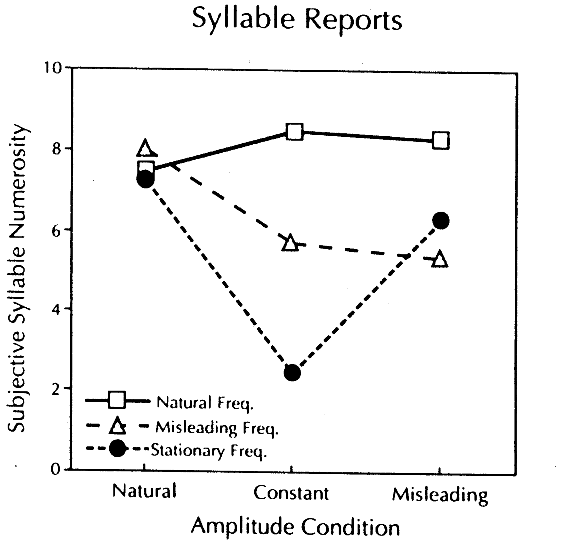


Figure 6. Average syllable numerosity reports in Experiment 2.

Post hoc tests (Newman-Keuls method of multiple comparisons, $\alpha = .05$) were used to assess the statistical differences among the nine means. To summarize the findings: First, accurate reports of syllable numerosity (eight syllables) were obtained in five conditions. Of course, this occurred in the case of natural frequency and amplitude presentation, but it was observed as well in the two cases of unnatural amplitude variation imposed on natural frequency variation, and in the complementary cases of unnatural frequency variation exhibiting a natural energy envelope. This shows that both frequency variation and amplitude variation can provide sufficient information for some aspects of syllable perception. Second, in the extreme case of constant tone frequency within an unchanging energy envelope, the syllable reports approached a mean of 2, perhaps suggesting that the intermittent tone following the two momentary occurrences of a fricative formant was the residual property of the displays on which listeners based their perceptual reports. Last, in the three conditions in which either amplitude or frequency properties, or both, were misleading, the reports were statistically identical, approaching a mean of 6 syllables.

Discussion

Experiment 2 can clarify the issues pertaining to the perception of tone analogues if we consider three aspects of the outcomes: (1) the parallels between the findings here and those in Experiment 1; (2) the circumstances in which frequency and amplitude properties influenced judgments of syllable numerosity; and (3) the surprisingly poor transcription performance that was observed in two of the three natural frequency conditions. Collectively, these points supplement the conclusion suggested by the first experiment, by suggesting a role of both frequency and amplitude variation in the perception of sinusoidal sentences.

First, in the four conditions that the two experiments shared, the pattern of syllable numerosity results in Experiment 2 was similar to the findings in Experiment 1. Briefly, in both experiments, there was little effect of misleading or constant amplitude properties on the apparent syllable pattern when the natural frequencies were also presented. Because this was observed here in a sentence of rather diverse phonetic composition, we can be relatively confident that the results of Experiment 1 were not attributable to the preponderance of continuant consonants used in that test. Again, it seems that sufficient acoustic information relevant to the perception of the pattern of syllables in a sine wave sentence is to be found within the properties of frequency variation of the component tones.

Second, this hypothesis must be supplemented to accommodate the obvious influence of amplitude patterns that were revealed in test conditions featuring natural amplitude values imposed on tones with misleading or constant-frequency properties. The availability of the natural amplitude pattern was adequate to guarantee accurate syllable reports despite the elimination of natural frequency variation, suggesting that natural amplitude properties, much like natural frequency patterns, are perceptually salient and adequate. This finding clearly supplements the conclusion suggested by Experiment 1, which emphasized solely the adequacy of information borne by frequency variation. Furthermore, the three conditions in which tone frequencies did not vary show the unmistakable perceptual effects of the amplitude manipulations on perception. Altogether, then, this test established the sufficiency of amplitude variation and replicated the proof of the adequacy of frequency variation in the perception of some attributes of syllables. A set of parametric evaluations of the perceptual sensitivity to natural signal variation may reveal the specific physical properties of the signal or the auditory mechanisms permitting natural variation to block the effects of distorted or unnatural forms of change. The present results show that such study is clearly warranted.

Third, a striking discrepancy is to be found in considering the transcription results of Experiments 1 and 2, in contrast to the close agreement in syllable counting. Unlike what occurred in the outcome of Experiment 1, here transcription performance deteriorated completely when anomalous amplitude patterns were used with natural frequency values. Because syllable reports were not similarly disrupted, this finding suggests that the phonetic impressions available to subjects in such conditions were prosodically related to the natural cases, but were unintelligible due to the obscurity of key phonetic features. Is this finding due to the imposition of unnatural amplitude contours that, by requiring tonal patterns to persist throughout the sentence, obscured the stop closures?

This speculation is encouraged by two kinds of evidence. The first is the precedent of Experiment 1, in which transcription performance was unaffected by the amplitude pattern imposed on the sinusoidal components. Perhaps this immunity from the impaired performance that was observed in Experiment 2 was due solely to the absence of stop consonants from that sentence. Certainly, one highly restricted hypothesis that may be based on the two experiments is that amplitude may vary freely without perceptual consequence if it is imposed on frequency variation that replicates the spectral properties of speech, and if there are few or no stop consonants in the sentence. The test of this hypothesis requires a replication in which unnatural amplitude variation occurs while the silences associated with stop consonants are conserved. The second kind of evidence encouraging this speculation is found in the general findings of sine wave replication. In contrast with the emphasis of some perceptual research that has highlighted the role of acoustic details in phoneme recognition—brief elementary cues, in other words—sine wave studies suggest that the perceiver need not be a meticulous listener attentive to a panoply of brief acoustic elements. The evidence of sinusoidal replication reveals the operation of a different kind of sensory susceptibility for which the coarse grain of the acoustic signal is effective. While it remains to be demonstrated conclusively that the sensitivity to time-varying signal attributes plays a crucial role in more acoustically ordinary cases (though, see Whalen, 1984, for a potential instance), one principle tentatively urged by this work is that perception of speech may rely much more on the coarse acoustic grain in frequency and amplitude than previous views have allowed. The crucial test of this requires the specification of the grain size of perceptual information in speech.

In the cases of amplitude variation in Experiment 2, we find that the silence during stop closures was a kind of gross property—the *complete* absence of energy—that may bear significant perceptual information. To evaluate this proposal, we conducted a third experiment to determine whether transcription performance was restored by the presence of appropriate silences in the signal, regardless of other attributes of the energy envelope that may be anomalous. This test called only for three amplitude conditions incorporating natural frequency variation, though unlike the second experiment, silences reflecting stop closures were preserved in misleading and constant amplitude test items. Our conclusion favoring gross energy changes would be warranted if the introduction of silences permitted the perception of segmental phonetic properties despite fine-grain anomalies in the energy envelope.

EXPERIMENT 3

In Experiment 2, we saw that the variations in the energy envelope had the greatest effect on perception of syllables when frequency variation was unnatural and perhaps neutralized as a source of phonetic information. In contrast with the effects on syllable reports, the effects on transcription—a fairly direct measure of segment perception—were most dramatic in the conditions presenting natural frequency variation. It was surprising, nonetheless, to see transcribability deteriorate so completely in the two energy conditions that departed from the natural values. Experiment 1 offered no clue that this might occur, perhaps because the phonetic variety and composition of the test sentence in Experiment 1 made it immune to any such effect. To resolve the difference between these two findings—one suggesting that an anomalous energy envelope does not affect transcription, the other that it hugely affects transcription—required a third test.

In Experiment 3 we evaluated the hypothesis that phonetically important information can be conveyed by gross changes in the energy envelope. Though this hypothesis is newly invoked for explaining the perception of spoken messages from sinusoidal vehicles, it is an established finding in linguistic phonetics (Halle et al., 1957; cf. Fitch, Halwes, Erickson, & Liberman, 1980). Applied to the case at hand, the hypothesis predicts that the sentence used in Experiment 2, which had stop consonants in crucial positions, was difficult to perceive because the silences correlated with vocal tract closure were eliminated.[2] It also permits the conclusion, based on the findings in Experiment 1, that an anomalous amplitude envelope may not disrupt segment perception when there are few or no stop consonants at stake.

Method

Subjects. Fifty-two subjects were tested. Each listener reported normal hearing in both ears, and none had participated in studies in which sine wave replicas of utterances had been employed. Some subjects received course credit, and others were paid for participating.

Acoustic test materials. Three sinusoidal sentence patterns based on the utterance "My t.v. has a twelve-inch screen" were used to compose the test. One, which had natural frequency and amplitude parameters, was identical to the pattern used in Experiment 2. The second and third patterns were derived, respectively, from the *natural frequency, misleading amplitude* test item and the *natural frequency, constant amplitude* test item used in Experiment 2. In both cases, the amplitude patterns that were imposed on the natural frequency variation remained misleading or constant, with the exception that silences occurring during consonant closures were restored by making the necessary changes to the sine wave synthesis parameters. Silence durations were adopted from the natural amplitude values.

Sinusoidal synthesis produced digital records of the waveforms, which were converted to analog signals and recorded on audiotape. Test signals were presented from tape, attenuated to approximately 65 dB SPL, over matched and calibrated headsets.

Procedure. A session began with an eight-sentence pretest, in the same manner as in Experiment 2, at the conclusion of which the critical test sentence was presented. Each sentence occurred eight times, separated by 3 sec, with 10 sec between trials. The subjects were asked to transcribe a sentence synthesized by a computer. There were three test conditions in this experiment, corresponding to the three different amplitude patterns imposed on the natural frequency values: (1) natural amplitude, (2) misleading amplitude with silent closures, and (3) constant amplitude with silent closures. The subjects were assigned randomly to conditions.

Results and Discussion

Seven subjects misunderstood the test instructions or transcribed none of the pretest sentences and were eliminated from the data set (7 of 52 = 13%), leaving 15 subjects in each of the three test conditions. The outcome of the test was quite clear: Transcription performance was good in all three conditions. Mean performance for the natural amplitude group was 5.2 syllables correct; for the misleading amplitude group it was 3.3 syllables; for the constant amplitude group it was 4.1 syllables. Two statistical tests determined that transcriptions differed from a hypothetical mean of 0 syllables correctly transcribed, but that performance did not differ across the three test conditions. First, the statistical difference between the grand mean of the data set and a hypothetical mean performance of 0 syllables correct was assessed by t test, which indicated a highly significant nonzero performance in these conditions [$t(44) = 10.42, p < .0005$]. Second, a one-way analysis of variance performed on these data revealed no significant differences in performance among the three test conditions [$F(2,42) = 1.896, p > .1$].

The conclusion prompted by these results fits well with those of our prior tests, pointing clearly to the phonetic perceptual importance of coarse-grain variation in signal amplitude. It resolves the discrepancy between the first and second experiments of this set, namely, in identifying the role of silence as a conspicuous acoustic marker of consonantal closures. By hindsight, we can see that the elimination of silent portions of the signal can only have had minimal effect in Experiment 1, inasmuch as the only segment affected was the single stop consonant in the test sentence. In Experiment 2, the deterioration of transcription performance was more pronounced due to the presence of stops throughout the sentence. Perceivers in the present test tolerated the departures from natural amplitude variation in fine detail, we may presume, because in coarse grain the perceptually critical closures were well evident.

GENERAL DISCUSSION

This set of three experiments was motivated by the need to assess the contribution of the sole acoustic aspect of sinusoidal sentence replicas that presents the listener a veridical aspect of the speech signal: the energy envelope. Sine wave replicas preserve the variation in amplitude of the individual vocal resonances, despite the fact that the rise and fall of the energy pattern is imposed on unnatural carriers. Although it had seemed exceedingly likely that frequency variation of the individual tones was the principal source of information in the perception of sentence analogues, the empirical support for this hypothesis was equivocal. In fact, the role of metrical properties in lexical information processing was well supported, and one plausible acoustic basis for this metrical information is the cyclical variation in the energy envelope, the same veridical attribute of speech preserved by sine wave sentences. An alternative to our initial hypothesis—that the listener attends to time-varying frequency information, instead of momentary, elementary speech cues—was therefore warranted. In contrast, we considered the hypothesis that the listener is unable to derive much from tone frequency variation after all, but is able to transcribe such signals by relying on the veridical properties of the energy envelope when the weird timbre, or sharp spectral peaks, or harmonically unrelated components of tone complexes yield ambiguous or defective phonetic information.

In fact, the picture drawn by the three experiments reported here differs from both of these characterizations. There is no denying now that both frequency and amplitude variation are capable of producing impressions of syllable variation, independently of each other in some circumstances. More germane to the case of sinusoidal replicas, though, is the finding that phonetic perception seems to depend crucially on the concurrent availability of natural frequency variation and gross amplitude variation. As far as we can tell from immediate evidence, the perceptually critical amplitude information was, grossly, the presence or absence of signal, a correlate of obstruction of the vocal tract. Other less extreme variations in amplitude level had no measurable effect in our tests, and on that evidence they are consigned to the class of the fine acoustic grain. Converging tests are still required to evaluate this, to be sure, but it seems after all that the perceptually important aspect of the amplitude variation is to mark the stop consonants.

What does this result say about ordinary speech perception? Considered broadly, this result is understandable from the perspective of studies on the robustness of spoken communication. It appears that little information may be carried by the amplitude pattern alone, for amplitude distortion is rarely devastating to segment intelligibility (Klatt, 1985; Licklider, 1946; Miller, 1946). Moreover, despite the temptation to treat the syllable as if it were an acoustic unit—because the energy envelope is correlated with the cyclical opening and closing of the vocal tract—the syllable is also a linguistic unit. To the cross-language evidence on the composition of the syllable (e.g., that of Price, 1980) and to the phonological arguments (e.g., those of Kiparsky, 1979) we can add this perceptual evidence about the specific value of amplitude variation in signaling stops rather than syllable trains.

Many of the parallels between the listener's perceptual treatment of natural signals and sinusoidal replication remain to be shown. However, the agreement between our data and those in other relevant studies of language encourage the generalization of our present finding—that syllable perception depended on the same acoustic properties as did segment perception—to the claim that the syllable is a multiply determined linguistic unit corresponding to no simple property of the acoustic signal. Collectively, the diverse evidence suggests that the perception of sinusoidally recoded signals is similar to the perception of natural speech, and this apparent congruence seems again to indicate that such tonal replicas preserve information ordinarily available in natural acoustic signals. The listener who contends with a sinusoidal sentence, in our view, makes use of this informative time-varying residue, and does little that is exceptional to perceive the linguistic properties. Nonetheless, in order to reveal the operation of perception from coarse-grain properties, it is essential to employ acoustic signals that leave the perceiver no alternative but to rely on nonelemental time-varying attributes.

To summarize, in three experiments, we have added to the evidence that the perception of sinusoidal replicas of speech signals is based on the coherent frequency variation of the tonal components. The first experiment in this report rules out the possible counterargument that perception occurs only through a process of attention to syllable properties first, as if this were possible on the acoustic basis of the veridical amplitude variation that the tones present. On the contrary, amplitude variation can count for little, and frequency variation for much. The second experiment showed that the perception of some attributes of syllables endured the elimination or distortion of natural frequency or energy variation. Last, the third experiment showed that the intelligibility of sine wave replicas depended on the concurrent availability of variation in frequency and in amplitude, though the fine grain of amplitude variation appeared to be negligible, perceptually, when the overall coarse properties were preserved. These findings permit us to extend the alternative to the familiar characterizations of the perception of speech based on discrete cues—that perception can be keyed to acoustic variation, independently of the specific fine-grain acoustic details that compose the signal—which we have derived from considering the phenomenon of tone analogues of speech.

REFERENCES

Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception & Performance*, **6**, 536-563.

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

Cutler, A., & Foss, D. J. (1977). On the role of sentence processing. *Language & Speech*, **20**, 1-10.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception & Performance*, **14**, 113-121.

Darwin, C. J., & Bethell-Fox, C. E. (1977). Pitch continuity and speech source attribution. *Journal of Experimental Psychology: Human Perception & Performance*, **3**, 665-672.

Egan, J. (1948). Articulation testing methods. *Laryngoscope*, **58**, 955-991.

Fant, C. G. M. (1962). Descriptive analysis of the acoustic aspects of speech. *Logos*, **5**, 3-17.

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics*, **27**, 343-350.

Fowler, C. A., & Smith, M. R. (1986). Speech perception as "vector analysis": An approach to the problems of invariance and segmentation. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 123-138). Hillsdale, NJ: Erlbaum.

Halle, M., Hughes, G. W., & Radley, J.-P. A. (1957). Acoustic properties of stop consonants. *Journal of the Acoustical Society of America*, **29**, 107-116.

Huggins, A. W. F. (1978). Speech timing and intelligibility. In J. Requin

(Ed.), *Attention and Performance VII* (pp. 279-297). Hillsdale, NJ: Erlbaum.

Jenkins, J. J., Strange, W., & Edman, T. R. (1983). Identification of vowels in "vowelless" syllables. *Perception & Psychophysics*, **34**, 441-450.

Kewley-Port, D., & Luce, P. A. (1984). Time-varying features of initial stop consonants in auditory running spectra: A first report. *Perception & Psychophysics*, **35**, 353-360.

Kiparsky, P. (1979). Metrical structure assignment is cyclic. *Linguistic Inquiry*, **10**, 421-441.

Klatt, D. H. (1985). A shift in formant frequencies is not the same as a shift in the center of gravity of a multiformant energy concentration. *Journal of the Acoustical Society of America*, **77**, S7.

Liberman, A. M., & Cooper, F. S. (1972). In search of the acoustic cues. In A. Valdman (Ed.), *Papers in linguistics and phonetics to the memory of Pierre Delattre* (pp. 329-338). The Hague: Mouton.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 421-461.

Licklider, J. C. R. (1946). Effects of amplitude distortion upon the intelligibility of speech. *Journal of the Acoustical Society of America*, **18**, 429-434.

Massaro, D. W. (1987). Categorical partition: A fuzzy logical model of categorization behavior. In S. Harnad (Ed.), *Categorical perception* (pp. 254-283). New York: Cambridge University Press.

Miller, G. A. (1946). Intelligibility of speech: Effects of distortion. In *Transmission and reception of sounds under combat conditions* (pp. 86-108). Washington, DC: National Defense Research Committee.

Nakatani, L. H., & Schaffer, J. A. (1978). Hearing "words" without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America*, **63**, 234-245.

Price, P. J. (1980). Sonority and syllabicity: Acoustic correlate of perception. *Phonetica*, **37**, 327-343.

Remez, R. E. (1987). Units of organization and analysis in the perception of speech. In M. E. H. Schouten (Ed.), *Psychophysics of speech perception* (pp. 419-432). Dordrecht: Martinus Nijhoff.

Remez, R. E., & Rubin, P. E. (1983). The stream of speech. *Scandinavian Journal of Psychology*, **24**, 63-66.

Remez, R. E., & Rubin, P. E. (1984). On the perception of intonation from sinusoidal sentences. *Perception & Psychophysics*, **35**, 429-440.

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception & Performance*, **13**, 40-61.

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

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives in the study of speech* (pp. 1-38). Hillsdale, NJ: Erlbaum.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics*, **35**, 49-64.

NOTES

[1] It is typical for linear prediction estimates of vocal resonances to contain frequency values that depart dramatically from the acoustic spectrum at points of rapid spectrum change. Such results are common at the onsets and offsets of silence associated with closures and releases of the vocal tract, as occur in instances of stop consonants. In order to correct the results of linear prediction analyses for use as synthesis parameters, it is therefore necessary to check the computed values against unbiased acoustic analyses, such as, for example, those produced by the sound spectrograph.

[2] An apparent stop closure without associated acoustic silence was observed by Darwin and Bethell-Fox (1977). In that instance, though, the impressions of the stop closure were also associated with a gross spectral change. But it was a gross change in the fundamental frequency of phonation, creating an impression of two talkers uttering speech in succession, rather than a silent closure within a single speech signal creating an impression of a closure within a single utterance.

(Manuscript received October 3, 1989;
revision accepted for publication June 5, 1990.)