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. Research has attempted to characterize this perceptual process by analyzing the acoustic properties of speech signals that specify the linguistic content (Fant, 1962; Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967; Mattingly, 1972; Stevens & Blumstein, 1978). 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 results suggest that the signal was also perceived, simultaneously, to be nonspeech. These novel findings imply that the process of speech perception makes use of time-varying acoustic properties that are more abstract than the characteristic spectra and speech cues typically studied in speech research.

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 talker, 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) (Markel & Gray, 1976). These values were hand-smoothed in some portions to ensure continuity, and were used as synthesis parameters for a digital sinewave 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 speech utterance. Figure 1 shows
Figure 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. Note that 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 there is no energy present except at the format center frequencies. The figure does not accurately reproduce the amplitude variation in the sinusoidal pattern.
narrowband and wideband spectrograms of the original spoken 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 (Chiba & Kajiyama, 1941; Fant, 1960) \[1\]. 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 \[2\]. Our tone stimuli consist of no more than three sinusoids, and therefore 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 literally no formant structure to the three-tone complexes, though 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 formant frequency variation. 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 finegrained 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 utterance. Overall, our three-tone pattern is a deliberately abstract representation of the time-varying spectral changes of the naturally produced utterance, though 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 assumed traditionally to underlie speech perception. None of the appropriate acoustic cues based on the acoustic events within speech signals is present in our stimuli, for example, neither formant frequency transitions, which cue
manner and place of articulation; nor steady state formants, which cue vowel color and consonant voicing; nor fundamental frequency changes, which cue voicing and stress (Liberman & Studdert-Kennedy, 1978). Similarly, the short-time spectral cues, which depend on precise amplitude and frequency characteristics across the harmonic spectrum, are absent from these tonal stimuli, for example, the onset spectra that are often claimed to underlie perception of place features (Stevens & Blumstein, in press). The perceptual importance of these attributes of speech signals has been rationalized 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 (Chiba & Kajiyama, 1941; Stevens, 1964). 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, these tests 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.

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 to be 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 [3]. Within each instructional condition, different groups of eighteen 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 twenty-one experimental conditions in all. In each condition a given sinusoidal pattern was presented four times in succession, at approximately 85 dB SPL, 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 in advance of 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 the sentence, "Where were you a year ago?" This outcome might be anticipated only if there were
# Table 1

Response Categories and Frequencies by Stimulus Condition in Instructional Condition A

<table>
<thead>
<tr>
<th>STIMULUS CONDITION</th>
<th>RESPONSE CATEGORIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 (T1+T2+T3)</td>
<td>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)</td>
</tr>
<tr>
<td>S2 (T1+T2)</td>
<td>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)</td>
</tr>
<tr>
<td>S3 (T2+T3)</td>
<td>Science fiction sounds (14), Radio interference (3), Music (2), Computer bleeps (2), Whistles (1), Several simultaneous sounds (1)</td>
</tr>
<tr>
<td>S4 (T1+T3)</td>
<td>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)</td>
</tr>
<tr>
<td>S5 (T1)</td>
<td>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)</td>
</tr>
<tr>
<td>S6 (T2)</td>
<td>Sirens (7), Bird sounds (6), Mechanical sound effects (4), Radio interference (4), Animal cries (3), Whistles (2), Computer bleeps (1)</td>
</tr>
<tr>
<td>S7 (T3)</td>
<td>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)</td>
</tr>
</tbody>
</table>
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 exact sentence is exceedingly small (Miller & Chomsky, 1960).

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 relative to the original utterance, "Where were you a year ago?" Average transcription performance in each stimulus condition is presented in Figure 2a. It is clear that a large number of subjects can identify the sentence in Conditions S1 and S2. Nine of the listeners across these two conditions transcribed the entire sentence correctly, though ten others reported that they could hear no sentence at all in the tones. The remaining listeners transcribed various syllables correctly. We conclude from these first two instructional conditions that naive 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, though not carried by acoustic elements producible by a vocal tract, is preserved in the time-varying relational structure of the stimulus pattern [4].

In Instructional Condition C, listeners were asked directly to evaluate the speech quality of the tone stimuli. They were told that they would be presented with the sentence, "Where were you a year ago?" and they were asked to make three judgments. First, they reported whether the sentence was discernible in the tonal pattern by responding Yes or No; they also provided a confidence rating for their judgments using a dual five-point scale. These responses were converted to a ten-point scale (1=confident Yes; 10=confident No). The scores are presented in Figure 2b grouped by stimulus condition. In five of the stimulus conditions, listeners were very confident that they did not hear the sentence in the tones. However, in Conditions S1 and S2, 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 Condition S2 (Scheffe 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 (1=all, 2=most, 3=a few, 4=almost none, 5=none). As shown in Figure 2c, for five of the stimulus conditions subjects indicated that they could not identify any of the words in the sentence. But, in the three-tone condition (S1), listeners reported that almost every word was clear. The omission of Tone 3 from the pattern in Condition S2 led subjects to report that significantly fewer words were intelligible (Scheffe test, p<.025), yet this condition remains significantly different from Conditions S3 through S7 (Scheffe test, p<.001).

In the third task, listeners rated the voice quality of the tone stimuli [1=natural, 2=funny (peculiar), 3=unnatural, 4=nonspeech]. The average ratings appear in Figure 2d. The split between S1 and S2 and the other conditions is still quite evident, as it was in Condition B above; however, we see here that these two stimulus patterns were judged to have unnatural voice quality despite their clear intelligibility. In essence, listeners apprehend
Figure 2. (a) Transcription performance for Instructional Condition B. (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, 2=most, 3=a few, 4=almost none, 5=none); (d) Naturalness ratings (1=natural, 2=peculiar, 3=unnatural, 4=nonspeech). Cross hatched=three-tone stimulus; hatched=two-tone stimulus; filled= single-tone stimulus.
the linguistic significance of the tonal patterns despite the radically unnatural, nonspeech quality [5,6]. 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 the present study cannot be explained within the framework of existing theories of speech perception [7], for the tones contained none of the elemental acoustic cues typically held to underlie speech perception (i.e., formant structure, fundamental period, or distinctive short-time spectra). Though the tones present 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 naive listeners in Condition A. 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 [8]. The linguistically primed listeners in Conditions B and C are 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 provide sufficient stimulation to evoke phonetic perception, albeit a kind that also identifies 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 sinewave spectra [9]. Further examples of nonspeech tonal analogues of natural speech utterances are needed to characterize more precisely the time-varying relations within the acoustic patterns that support phonetic perception.

REFERENCES

Best, C. T., Morrongiello, B., & Robson, R. The perceptual equivalence of two acoustic cues for a speech contrast is specific to phonetic perception. Perception & Psychophysics, in press.
Fant, C. G. M. Descriptive analysis of the acoustic aspects of speech. Logos, 1962, 5, 3-17.


FOOTNOTES

1. The closely spaced horizontal lines shown in Figure 1a are the harmonics of the fundamental frequency of phonation, and are typically revealed in narrowband spectrograms.

2. Typically, the amplitude of the valleys in the spectrum of natural speech ranges from 10-30 dB below the amplitude of the peaks (Stevens & Blumstein, in press).
Our listeners were students of introductory psychology at Indiana University in Bloomington. They were naive with respect to synthetic speech.

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 intelligibility (Liberman & Cooper, 1972). For this reason, listeners probably do not require stimuli to display the acoustic "stigmata" of speech to be candidates for phonetic interpretation (Liberman, Mattingly, & Turvey, 1972). However, even schematized synthetic speech has consisted of acoustic cues that are utterable in principle as components of a speech signal; these cues enjoy 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 can neither be explained by arguing that they are components of natural signals; nor by arguing that they are acoustic products of vocal articulation.

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—which affect speech quality nonetheless (Liberman & Cooper, 1972). 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 (Allen, 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.

Although the intelligibility of our sinusoidal sentence is predicted by the co-occurrence of T1 and T2, but not of T1 and T3, the effectiveness of each tone pair will vary as a function of the phonetic composition of the utterance. While the resonance associated with the oral cavity is primary in its importance for phonetic perception (Kuhn, 1979), either F2 or F3 may be affiliated with the oral cavity, depending on the phone in question (Stevens, 1972). Therefore, the critical tone pair will sometimes include T2, sometimes T3, depending on the phonetic composition of the utterance.

The proposal that listeners "track" formant frequency variations must 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 the concept of the formant refers to a natural resonance of the vocal cavities (Hermann, 1894). Quite literally, then, there are no vocal resonances in our tone complexes (though listeners who succeed in extracting the meaning of the "utterance" 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 tone signals is that the former contain broadband formant structure and harmonic spectra, and the latter merely inharmonic peaks with infinitely narrow bandwidths.
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 (Köhler, 1910; Modell & Rich, 1915; Titchener [described in Boring, p. 374, 1942]). 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 low uncertainty judgment tasks (Cutting, 1974; Bailey, Summerfield & Dorman, 1977; Best, Morrongiello & Robson, in press; Grunke & Pisoni, 1979). The present study, however, makes use of neither a closed response set nor a low uncertainty task to obtain the effect of intelligibility.

We have recently synthesized the sentence, "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.