*Perception & Psychophysics*
*1984, 35 (5), 429-440*

462

# On the perception of intonation from sinusoidal sentences

ROBERT E. REMEZ
*Barnard College of Columbia University, New York, New York*

and

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

Listeners can perceive the phonetic value of sinusoidal imitations of speech. These tonal replicas are made by setting time-varying sinusoids equal in frequency and amplitude to the computed peaks of the first three formants of natural utterances. Like formant frequencies, the three sinusoids composing the tonal signal are not necessarily related harmonically, and therefore are unlikely to possess a common fundamental frequency. Moreover, none of the tones falls within the frequency range typical of the fundamental frequency of phonation of the natural utterances upon which sinusoidal signals are based. Naive subjects nevertheless report that intelligible tonal replicas of sentences exhibit unusual "vocal" pitch variation, or intonation. The present study attempted to determine the acoustic basis for this apparent intonation of sinusoidal signals by employing several tests of perceived similarity. Listeners judged the tone corresponding to the first formant to be more like the intonation pattern of a sinusoidal sentence than: (1) a tone corresponding to the second or third formant; (2) a tone presenting the computed missing fundamental of the three tones; or (3) a tone following a plausible fundamental frequency contour generated from the amplitude envelope of the signal. Additionally, the tone reproducing the first formant pattern was responsible for apparent intonation, even when it occurred in conjunction with a lower tone representing the fundamental frequency pattern of the natural utterance on which the replica was modeled. The effects were not contingent on relative tone amplitude within the sentence replica. The case of sinusoidal sentence "pitch" resembles the phenomenon of *dominance*, that is, the general salience of waveform periodicity in the region of 400-1000 Hz for perception of the pitch of complex signals.

A number of recent studies of speech perception have examined the effects of sinusoidal replication of speech signals (Bailey, Summerfield, & Dorman, 1977; Best, Morrongiello, & Robson, 1981; Grunke & Pisoni, 1982; Schwab, 1981). Typically, tonal analogs of speech are composed of three time-varying sinusoids, each tone reproducing the frequency and amplitude variation, sometimes schematically, of a formant from a natural utterance. In such acoustic patterns, devoid of the harmonic series and broadband formant structure of natural speech, the short-time acoustic properties are unmistakably not speechlike. Both acoustically and perceptually, sinusoidal signals are grossly unnatural, and naive listeners tend therefore to perceive sinusoidal sentences merely as several covarying tones unless they expect to hear a linguistic message; moreover, phonetic perception fails to occur unless the tonal stimulus is adequately structured, indicating that an explanation of this effect should be sought in terms of the information provided by these atypical stimuli (Remez, Rubin, Pisoni, & Carrell, 1981). When sinusoidal patterns are perceived phonetically, they are judged to be intelligible yet unspeechlike, presumably because they convey segmental information in an abstract pattern of spectrum *variation*, with almost none of the typical acoustic details of natural speech.

One consequence of this finding is methodological. This technique for transforming the signal can be used to reveal the perceptual significance of time variation in the speech stream. This is so precisely because such unspeechlike signals disentangle the pattern of frequency variation over time in the speech stream from the sequence of particular momentary acoustic elements that are produced by vocal articulation. In view of the acoustic differences between sinusoidal signals and the natural utterances that they replicate, it seems fair to suppose that sinusoidal replication does not merely reduce the amount of information present in the signal, as minimal-cue speech synthesis

The authors thank Peter Balsam, Louis Goldstein, and Leigh Lisker for advice and encouragement. We are also indebted to our reviewers for insisting that we include Experiment 4 in the report. This work was supported by Grants HD-15672 (to R.E.R.) and HD-01994 (to Haskins Laboratories). Address correspondence to: R. E. Remez, Department of Psychology, Barnard College, 606 West 120th Street, New York, NY 10027.

does (e.g., Delattre, Liberman, & Cooper, 1955; Abramson & Lisker, 1965). In that technique, a subset of the acoustic ingredients of an utterance is selected for imitating synthetically. Obviously, the information provided by natural acoustic elements is lost if those elements fail to appear in the synthetic replica. In such circumstances, phonetic information may or may not be adequately conveyed by the remaining acoustic structure. Therefore, this minimalist method is designed to reveal the effectiveness of particular acoustic elements—for example, a burst of noise, a low-frequency murmur, or a proscribed frequency transition in the second formant—when others have been neutralized or eliminated.

In contrast, the transformation of a speech signal into time-varying sinusoids does not preserve particular constituents of the acoustic signal while discarding others. Rather, it destroys the physical similarity of acoustic moments in natural speech and those in the sinusoidal patterns. The residual similarity between speech and sinusoidal imitations is to be found only in the variation of the two kinds of signal, and specifically in the pattern of frequency variation over time. For this reason, a significant aspect of the sinusoidal replication technique would be obscured by classifying the signals simply as "impoverished stimuli." They are, in fact, literal imitations of the time-varying properties of the supralaryngeal vocal-tract resonances. Sinusoidal signals of this type present the pattern of resonance center-frequency variation through an utterance, although the signals obviously do not contain formant structure.[1] Our tests (Remez et al., 1981) have established the sufficiency of this acoustic abstraction of the speech signal, in contrast to research which more customarily demonstrates the perceptual uses of selected brief pieces of the signal. When perceivers detect phonetic structure in sinusoidal patterns, this reveals the usefulness of the forms of stimulus change as phonetic information, and the independence of perception from most of the specific acoustic details with which the forms are conveyed.

### Sinusoidal Intonation

In an obvious way, however, sinusoidal replicas of speech are impoverished, despite all. The principal perceptual correlate of sentence intonation, the fundamental frequency of phonation (Lieberman, 1967), is absent from sinusoidal signals, which imitate only the frequency variation of the formant peaks. As a result of this deficiency, listeners have consistently reported that sinusoidal sentences exhibit noticeably weird patterns of intonation.[2] The perception of relative syllable stress (Fry, 1958; Lehiste & Peterson, 1959; Morton & Jassem, 1965), or of the placement of clause boundaries (Collier & t'Hart, 1975; Lehiste, 1973; Streeter, 1978), each of which is said to follow occasionally from normal intonation, must therefore be supported (if at all) by other means, because the anomalous intonation of sinusoidal replicas is quite different from the normal intonation patterns to which these roles are attributed. To the same extent that the fundamental frequency of an utterance also contributes segmental information [about consonant voicing (Summerfield & Haggard, 1977) or vowel identity (House & Fairbanks, 1953), for example], the listener will also be forced to rely on other, alternative sources.

But why do sinusoidal signals create this impression of peculiar intonation in the first place? Prosodic perception is an admittedly complex affair, in which the properties of a single piece of the acoustic stream may affect the recognition of segmental, syllabic, and syntactic structural properties together. In the sinusoidal case, it seems that the pattern of tones imitating only the formant variation inadvertently presents an effective stimulus for perceiving intonation. It is far from obvious why three tones in the frequency range of formants should lead to this impression of vocal pitch, for the acoustic properties corresponding to intonation typically occur several octaves below the lowest formant, and, consequently, below the lowest frequency tone in our three-tone patterns. We undertook the present study to identify the acoustic and perceptual basis for this peculiar concomitant of phonetic perception with sinusoidal signals. The first experiment described here determmined which of the likely acoustic sources for the anomalous intonation would, in fact, be identified as the correlate of sinusoidal intonation. The second experiment tested the salience of the empirically determined acoustic correlate of sinusoidal intonation, the tone reproducing the pattern of the first formant (Tone 1), as a function of its relative amplitude in the three-tone pattern. The third experiment revealed that subjects did not hear the intonation of a sinusoidal sentence as the correlate of Tone 1 when that tone was removed from the sinusoidal sentence pattern. Finally, the fourth experiment that we describe found that the intonation of a four-tone pattern, composed of three sinusoids imitating formant variation and a fourth imitating fundamental frequency variation, was again correlated with the first formant tone and not with the lowest frequency tone of the pattern, complementing the results of the prior three studies.

## EXPERIMENT 1

From the outset, there seemed to be at least three potential causes of the perceptual impression that sinusoidal replicas of natural utterances possess odd intonation. First, the apparent speech melody may be the listener's invention, given that the structure of the sinusoidal signal is defective precisely in representing the fundamental frequency of the original

utterance. Typical synthetic speech, on the other hand, is generated with a fundamental frequency pattern as well as a sequence of spectrum envelopes approximating the natural case. In the sinusoidal instance, the listener may fabricate an intonation pattern from the variation in the amplitude envelope of the signal, which is correlated with variation in fundamental frequency in the natural case (Lieberman, 1967), and which also is represented faithfully in sinusoidal replications of natural utterances.

Second, the listener may induce a pitch contour based on whatever changing harmonic relationships exist among the three tones of the sinusoidal pattern. The three tones are not likely to be related harmonically at any given instant, because they follow the computed resonance peaks and not the frequencies of the harmonics of the fundamental closest to the formant centers. Nonetheless, there may be a kind of auditory induction occurring, based on the varying relation of the frequencies of the three simultaneous tones, that produces a time-varying residue heard as the intonation contour. This possibility would be similar to the induction of the missing fundamental (Licklider, 1956; Schouten, 1940).

Third, the listener may use one of the three tones both for segmental information and for intonation information. Although the principal acoustic correlate of sentence intonation is the fundamental frequency, and although the fundamental frequency is present in the speech spectrum at an average of two octaves below the first formant, psychophysical and electrophysiological evidence suggests that listeners may detect the fundamental frequency of natural utterances by attending to the periodicity of the harmonics of the fundamental in the vicinity of the first formant (Greenberg, 1980). If an extrapolation of those findings is appropriate to the sinusoidal case, we would suspect the apparent intonation to be based on the pitch of the tone replicating the first formant of the natural utterance on which it was modeled.

To determine the basis for the apparent intonation of sinusoidal sentences, we performed a test of the apparent similarity of pitch contours, in which subjects judged one member of a pair of tone patterns as more like the speech melody of a sinusoidal sentence. The set of candidate intonation patterns included each of the three tones of the sinusoidal sentence pattern presented individually, a plausible fundamental frequency pattern derived from the amplitude envelope of the sinusoidal sentence, and a tone that reproduced the pattern derived by computing the greatest common divisor of the three tones at intervals of 10 msec throughout the sentence. On each trial, the subject was asked to identify the sentence melody of a three-component sinusoidal sentence presented once, and then to select the single-tone pattern from the two alternatives that was more like the melody of the sentence.

## Method

**Subjects.** Fifteen adults with normal hearing in both ears were recruited by handbill from the population of Barnard and Columbia colleges. All were native speakers of English, and none had participated in other experiments employing sinusoidal signals. The subjects were paid for their services.

**Stimuli.** The acoustic materials used in this test consisted of six sinusoidal patterns, one three-tone sentence pattern, and five single-tone patterns, produced by the sine-wave synthesizer at Haskins Laboratories. This software synthesizer generates sinusoidal patterns defined by parameters of frequency and amplitude for each tone, updated at the rate of 10 msec per parameter frame. The initial synthesis parameters were obtained by analyzing a natural utterance, the sentence "Where were you a year ago?" produced by one of the authors. This utterance was recorded on audiotape in a sound-attenuating chamber and converted to digital records by a VAX 11/780-based pulse-code modulation system using a 5-kHz low-pass filter on input and a sampling rate of 10 kHz. At 10-msec intervals, center-frequency and amplitude values were determined for each of the three lowest formants in this utterance by the analysis technique of linear prediction. In turn, these values were used as sine-wave synthesis parameters after correcting the errors that linear prediction is prone to commit. Generally, inappropriate values are easy to identify in the parameter table. They are likeliest to be found when the formant extraction routines are unable to identify any amplitude peaks in the spectrum, for example, when amplitude is low due to consonant closures. Formant patterns are also corrected if the analysis designates an extraneous "formant," which displaces the proper values to the next highest or lowest formant, for example, during rapid spectrum change. A full description of sinusoidal replication of natural speech is provided by Remez, Rubin, and Carrell (1981).

The sentence pattern that was matched to the natural utterance was composed of three time-varying sinusoids. Tone 1 corresponded to the first formant, Tone 2 to the second, and Tone 3 to the third. A Fourier spectrum for a section of the three-tone pattern is shown in Figure 1a. Note that the relative energy in the three tones decreases with increasing frequency, imitating the natural case, but that the broadband formant and harmonic structure common to voiced speech is not present. The five alternative single-tone patterns that were used to compose the pairs of alternatives were: Tone 1, Tone 2, or Tone 3, each a component of the sentence pattern that the subject heard at the beginning of each trial; a plausible fundamental frequency pattern (PFO) computed from the amplitude envelope; and the pattern comprising the values of the greatest common divisors (GCD) of the three concurrently varying tones in the replica of the natural utterance, computed for each 10-msec frame of the sinusoidal synthesis parameters. Each of the single-tone alternatives was produced with equal average power. The plausible fundamental pattern was derived by modulating the frequency of a 100-Hz tone to follow the changes in amplitude of the waveform of the sinusoidal sentence. The maximum range of this tone was 20 Hz, and the maximum rate of frequency change was 1 Hz/10 msec. Finally, the frequency values for synthesizing the "missing fundamental" tone were determined by computing the integer, for each synthesis frame, of greatest value that served as a divisor for each tone frequency, with no more than a 2% remainder. The average frequency value of this plausible missing fundamental tone was 92 Hz, well within the fundamental range of the talker producing the original utterance from which these six tonal patterns were derived. The amplitude values of this tone were matched for each 10-msec frame to the amplitude values of Tone 1. A graphic representation of each of the five single-tone patterns is presented in Figure 1b.

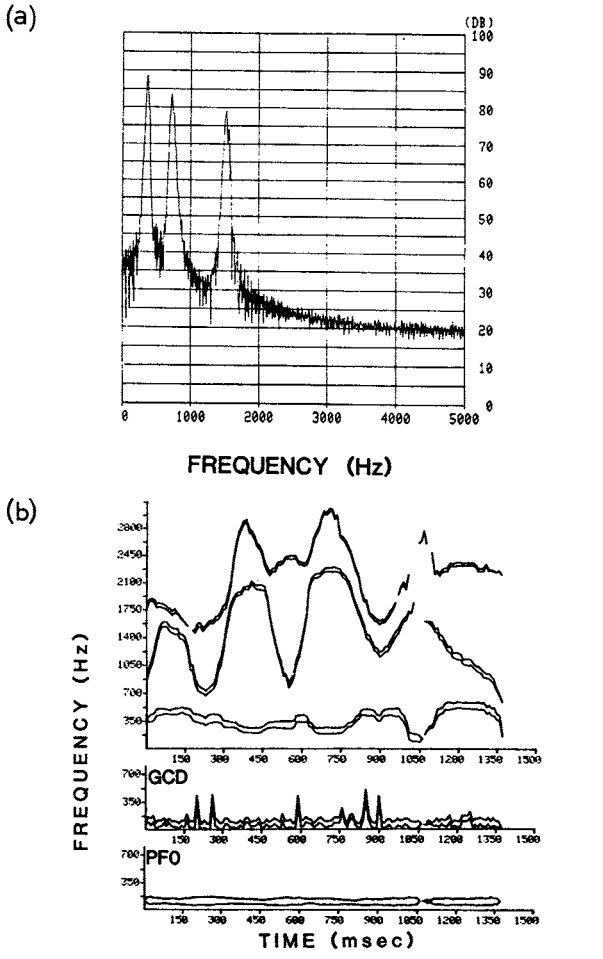


**Figure 1. (a) This panel is the Fourier spectrum of a representative section through the three-tone pattern replicating the sentence "Where were you a year ago?" (b) The five tone patterns used as stimuli in Experiment 1. Top panel: The three-tone pattern replicating the first three formant center-frequency values of the sentence "Where were you a year ago?" Middle panel: The pattern composed of the greatest common divisors (GCD) of the three tones in the sentence replica, computed for each 10-msec synthesis frame. Bottom panel: A plausible fundamental frequency pattern (PFO), computed from the amplitude envelope of the sentence pattern. In all cases, variation in thickness represents amplitude variation.**

The synthesized test materials were converted from digital records to analog signals, recorded on audiotape at Haskins Laboratories, and were presented to listeners in the Perception Laboratory of the Department of Psychology, Barnard College, by playback of the audiotape. Average signal levels were set at 72 dB SPL. Stimuli were delivered binaurally in an acoustically shielded room over Telephonics TDH-39 headsets.

**Procedure.** Listeners were told that the experiment was examining the identifiability of vocal pitch, the tune-like quality, of synthetic sentences. To illustrate the independence of phonetic structure and sentence melody, the experimenter sang the phrases "My Country 'Tis of Thee" and "I Could Have Danced All Night" with the original melodies and with the melodies interposed. When subjects acknowledged their ability to determine the melody of a sentence regardless of the words, they were instructed to attend on each test trial to the pitch changes of the sinusoidal sentence, to identify the pattern, and to select the alternative of the two following patterns that more closely resembled the pitch of the sentence. The subjects recorded their choices in specially prepared response booklets.

Each trial had the same format. First, the sinusoidal sentence "Where were you a year ago?" was presented once. Then one of the five single-tone patterns was presented. Finally, a second single-tone pattern was presented. There were 10 different comparisons among the five different single-tone alternatives. Counterbalanced for order, each subject judged each different comparison 10 times. Each sinusoidal pattern was approximately 1,400 msec in duration, the interval between items within a trial was 1 sec, and the silent interval between trials was 3 sec.

## Results and Discussion

Differences among the means of subjects' performance in the differential similarity test were identified with the analysis of variance. Irrespective of the order of alternatives within a trial, there were 10 different types of trial comparing tonal alternatives: Tone 1 vs. Tone 2, Tone 1 vs. Tone 3, Tone 1 vs. PFO, Tone 1 vs. GCD, Tone 2 vs. Tone 3, Tone 2 vs. PFO, Tone 2 vs. GCD, Tone 3 vs. PFO, Tone 3 vs. GCD, and PFO vs. GCD. For each type of trial, a signed value indexing the preference for one alternative or the other was computed by taking the difference of the number of trials (out of 10) on which the subject selected the first alternative versus the second. (The order of alternatives used to determine the sign of the difference was the order of the alternatives given directly above.) Note that if the subject had no consistent preference within a trial type, the index value approached 0; a consistent preference approached ±10. The one-way analysis of variance revealed a significant difference among the means of the similarity scores on different trial types [$F(9,126) = 11.8$, $p < .001$]. Scheffé post hoc means tests showed that Tone 1 was preferred to every alternative with which it was compared, but that in trials excluding Tone 1 the greatest performance difference was not significant. Histograms of the group data are shown in Figure 2. The figure represents the proportion of trials on which each alternative in each comparison type was selected. From this figure, it seems clear that the tone replicating the first formant is chosen as the sentence pitch in any comparison that includes it, and that in every other case the choice of tone is equivocal.

This outcome encourages a few conclusions about the cause of the odd intonation of sinusoidal sentences. It seems that the tone that replicates the first formant of the natural utterance is put to two uses, perceptually, by listeners. Although it seems to provide segmental information about consonants and vowels, as we expected, it also is serving as the acoustic correlate of sentence pitch, a function usually attributed to the fundamental frequency of phonation. This outcome seems surprising because Tone 1 in sinusoidal sentences is typically one and one-half octaves higher than the fundamental, as is the first for-

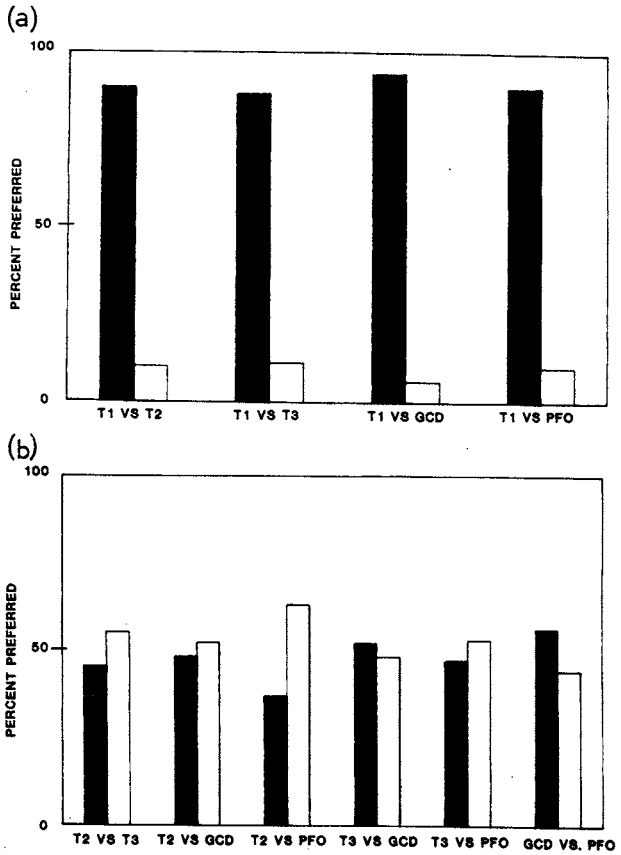


**Figure 2. Differential similarity data from Experiment 1. (a) Comparisons in which Tone 1 was an alternative. (b) Remaining comparisons.**

mant in natural utterances. Moreover, Tone 1 would be quite beyond the comfortable phonatory range of adults capable of producing the associated formant frequencies. The perceptual preference for Tone 1 as the intonation of the sentences is not to be expected, therefore, if perception is based primarily on the listener's knowledge of the normative articulatory abilities of talkers. Although some evidence implies that the variation in fundamental frequency in natural speech is correlated with formant pattern (House & Fairbanks, 1953; Lehiste & Peterson, 1961), no current proposal also suggests that the perceiver uses the first formant frequency variation as information both for intonation and for segment identification. This, however, seems to have occurred in the case of sinusoids.

Research on the phenomenon of the *dominance region* (e.g., Plomp, 1967; Ritsma, 1967; see also Greenberg, 1980) may begin to explain this result. These studies established that the impression of pitch corresponds to the shared fundamental period of the third through fifth harmonics, and not to the periodicity of excitation occurring in the lower or higher frequencies. In the nonspeech case (Plomp, 1967), listeners judged the apparent pitch of complex signals composed simultaneously of two different harmonic series. Each series presumably could have led to the impression of a different pitch, but the series falling within the "dominant" region in fact determined the pitch. In the speech case, Greenberg (1980) recorded evoked potentials to synthetic vowels in human subjects. He found that the auditory representation of fundamental frequency was strongest when the first formant occurred within the dominant region. If the impression of pitch is obtained from analysis of this band in the auditory representation, then the implication of this work is clear: A person listening to speech normally uses the region of the spectrum associated with the first formant to obtain periodicity information as well as to detect the frequency of the first formant itself. Ordinarily, the periodicity of the stimulus in this region and the frequency of the first formant will differ, although in the present case they happen to be identical.

We cannot be sure, though, that Tone 1 is selected for its prosodic role for any other reason than that it is the loudest tone in the three-tone complex. Recall that the parameters that are specified for each time-varying sinusoid in the replication of the natural utterance include a formant center-frequency and a formant amplitude specification, both derived by linear prediction analysis of the speech waveform. Because the first formant in natural speech commonly has the greatest power and each higher formant less energy, this spectrum envelope rolloff is preserved in the sinusoidal imitation. To identify the relation between the selection of Tone 1 as the pitch contour of the sinusoidal sentence and its relatively great acoustic power, we performed Experiment 2. In addition, we attempted to test the generality of our finding by using a new sentence.

## EXPERIMENT 2

In this portion of our study, we varied the relative amplitudes of the three tones composing the sinusoidal sentence and again employed a test of differential similarity to determine the alternative most similar to the intonation of the sinusoidal sentence. If, in Experiment 1, Tone 1 was judged most similar in pitch pattern to the intonation of the sinusoidal sentence merely because Tone 1 had the greatest power of the three components of the sentence pattern, then this should not recur when the relative amplitude differences of the three tones are eliminated, or reversed. On the other hand, if Tone 1 is the stimulus for intonation because it occurs within the dominance region, then we should not expect amplitude variation to change the differential similarity performance, as long as Tone 1 is detectable (Ritsma, 1967). This experiment, therefore, estimated the effects on apparent intonation of equating the amplitudes, and of inverting the order of amplitudes, among the tones of a three-component replica of a natural utterance. In

addition, we also employed a different sentence in the study, to identify any effects that may have been particular to the phonetic properties of the sentence used in the prior experiment.

## Method

**Subjects.** Fourteen listeners were drawn from the local population of audiologically normal undergraduates, again. None had been tested previously in studies of sinusoidal synthesis. The subjects received pay for participating.

**Stimuli.** A three-tone replica was prepared for the sentence "I read a book today," according to the procedure described in Experiment 1. From this replica, two versions were subsequently made. In the first, the tone amplitudes were set to be equal; in the second, the amplitude order was the inverse of the natural case, with Tone 3 possessing the greatest power and Tone 1 the least. Figure 3a shows the pattern of three tones composing the sentence. Figures 3b and 3c show Fourier spectra of sections of the equal (flat) amplitude and uptilted amplitude versions of this sentence.

The single-tone alternative patterns to be compared with the apparent intonation of the sinusoidal sentence on each trial consisted this time simply of each of the three individual tones composing the sentence. The single-tone alternatives were prepared as in Experiment 1, with equal average power. On each trial, the subject heard one of the two versions of the sentence, with the flat or the uptilted spectrum, followed by two of the three alternative tone patterns.

**Procedure.** Each trial began with a single presentation of the sinusoidal sentence "I read a book today," either the flat spectrum or the uptilted spectrum version. Following the presentation of the sentence were two single-tone alternatives, from which the subject selected the better match to the apparent intonation of the sentence. Collapsing over the counterbalancing of order for each pair of alternatives, there were three different types of trials: the comparisons of Tones 1 and 2, Tones 1 and 3, and Tones 2 and 3. Each of these was presented 20 times, 10 times in each order. In addition, 12 trials were interspersed in the test order in which a normal spectrum relationship occurred among the tones of the sentence, although the overall power was greatly reduced. With this quiet, normal amplitude rolloff sentence, the only alternative tonal intonation patterns presented were Tones 1 and 2. The data from this condition served as a converging check on the outcome of the prior investigation.

One hundred and thirty-two trials were presented in this test. On each trial, the subject identified the intonation of the sinusoidal sentence presented first and then selected the more similar of the two lagging alternative tone patterns. The choice was recorded in pencil in a specially prepared response booklet. There were 1 sec between items within a trial, 3 sec between trials, and 8 sec following every 12th trial.

## Results and Discussion

The judgments were handled in a manner analogous to Experiment 1. Signed preference scores were determined for the three comparisons of the flat and uptilted sentence conditions. For each comparison, the difference was computed between the number of trials on which the first alternative was chosen and the number of trials on which the second was chosen. In the computation of the difference scores, the alternatives were compared in this order: Tone 1 vs. Tone 2, Tone 1 vs. Tone 3, Tone 2 vs. Tone 3. A two-way repeated measures analysis of variance, with the factors of sentence (flat vs. uptilt) and comparisons (Tones 1 vs. 2, 1 vs. 3, and 2 vs. 3), was used

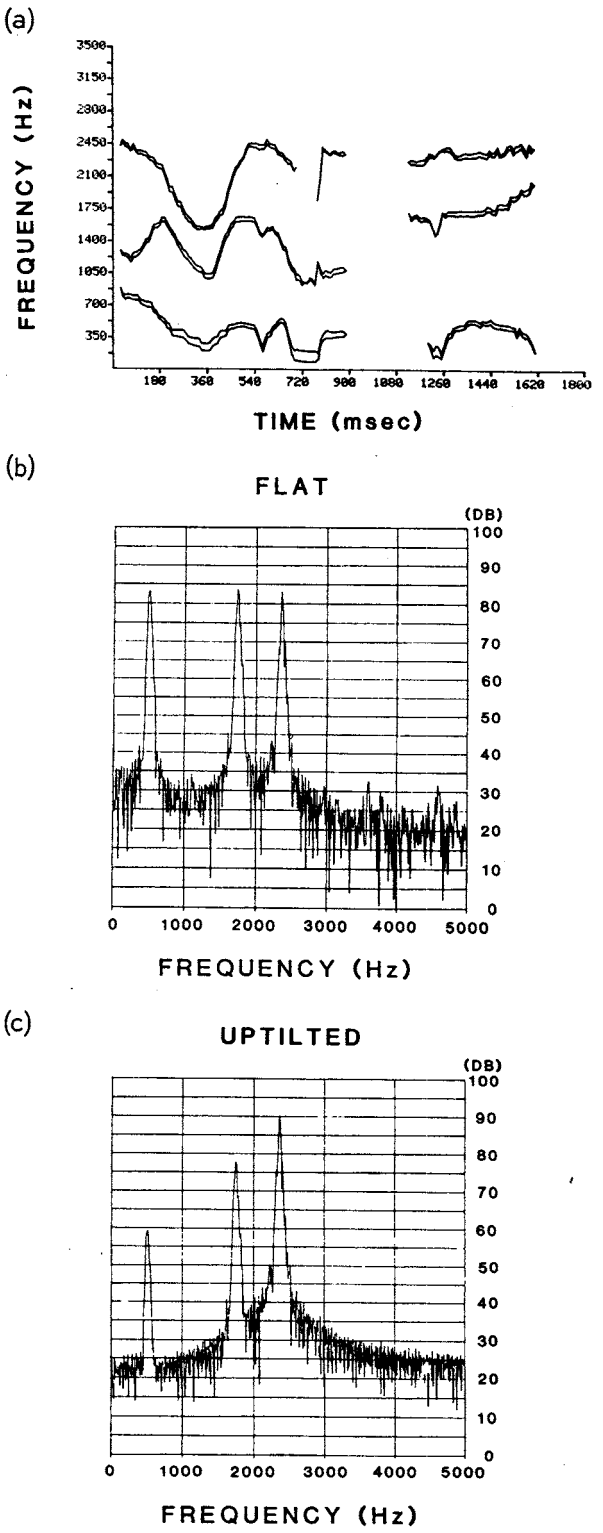


**Figure 3. (a) The three-tone pattern replicating the sentence "I read a book today." (b) Fourier analysis of a section through the flat spectrum version of the sentence (equal energy in each tone). (c) Fourier analysis of a section through the uptilted spectrum version. Compare with Figure 1a.**

to determine whether there was an effect of tone amplitude in the sentence on the perception of intonation. The data from the quiet normal amplitude trials, in which Tone 1 was the clear preference, were omitted from this analysis.

The group data are shown in Figure 4. It is obvious from that figure that Tone 1 retains its preferred status. This is confirmed by the analysis of variance. There was a main effect of sentence type, indicating that the preference scores were more consistent for the flat than for the uptilted sentences [$F(1,13) = 9.5$, $p < .01$]; and, there was a main effect of trial type [$F(2,26) = 10.1$, $p < .001$], with Tone 1 preferred in each of the two pairs in which it occurred, and no consistent preference between Tones 2 and 3. The interaction term was not significant [$F(2,26) = 0.6$, $p > .1$], indicating that the subjects preferred Tone 1 as the best match for sentence intonation regardless of the spectrum manipulation.

This experiment encourages the conclusion of our initial study of sinusoidal intonation. It seems that the functions of Tone 1 include both the segmental use typically associated with the first formant that it replicates and the use typically identified with the fundamental frequency of phonation in natural speech. The durability of the listeners' reliance on Tone 1 for intonation information is noteworthy, especially considering the inversion of the order of relative amplitudes among the tonal components of the signal. It suggests that the dual use of Tone 1 in sinusoidal sentences is brought about by virtue of its occurrence within the dominance region, and not because it is the component with greatest power. Periodicity within this frequency band, including instances of relatively low power, evidently determines the pitch pattern of the perceived sentence. It seems, then, that Tone 1 is concurrently represented as an amplitude peak in the spectrum, which provides information about segmental phonetic properties of the utterance, and also as a periodic spectrum element that determines the apparent pitch of the tonal complex. Ordinarily, in speech, the frequencies occurring within this region are harmonics of the fundamental frequency of phonation. However, in this anomalous case of formant center frequencies without harmonic excitation, there is no stimulation, periodic or otherwise, in the range of a talker's fundamental, and therefore no harmonics in the dominance region. There is, simply, the time-varying frequency of the tone following the formant, which is treated as the stimulus for pitch by default, regardless of its amplitude relative to the other components.

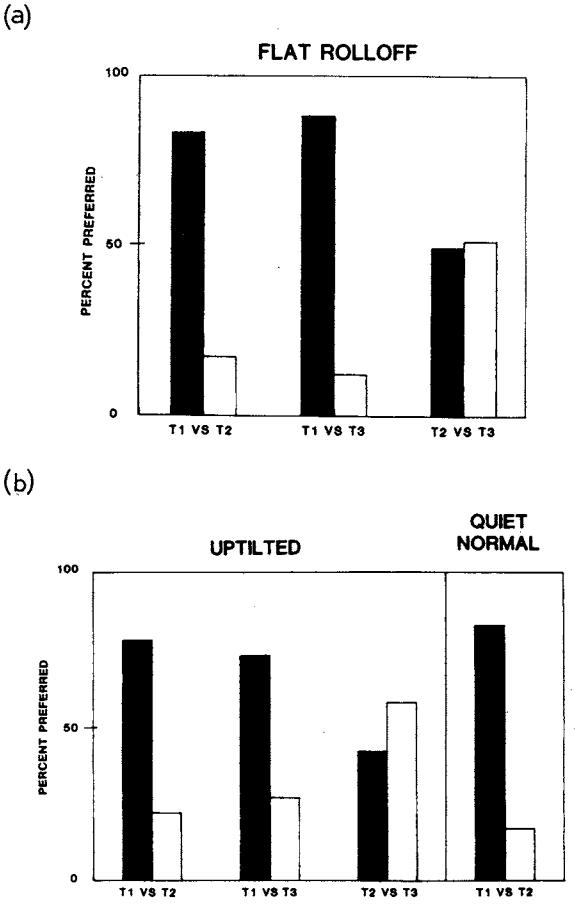


**Figure 4. Results of Experiment 2. (a) Differential similarity data for flat rolloff condition. (b) Differential similarity data for the uptilted and quiet-normal spectrum conditions.**

To conclude that the intonation of a sinusoidal replica is the correlate of Tone 1, and that this is attributable primarily to the occurrence of this time-varying periodic tone within the dominance region of the auditory system, we must establish that listeners reject Tone 1 as the best match of sinusoidal sentence intonation when the sentence does not include that tone. In other words, if a two-tone pattern, including only Tones 2 and 3, is presented in the same paradigm as in Experiments 1 and 2, listeners should *not* report that Tone 1 matches the intonation of this pattern. Were they to persist in identifying Tone 1 as the intonation pattern, we would be forced to conclude that the phenomenon of sinusoidal intonation was less a matter of the ordinary perception of extraordinary signals, as we have alleged, and was actually a matter of special induction of ad hoc attributes of an unfamiliar stimulus. Experiment 3 was performed to test whether Tone 1 was identified as the correlate of intonation for patterns that did not contain it.

## EXPERIMENT 3

At this point, the evidence shows that the tone following the frequency variation of the first formant is the correlate of the intonation of sinusoidal sentences. In all cases, Tone 1, corresponding to the track of the first formant, was judged more like the

sentence intonation than any other candidate. Our conclusion has emphasized the listeners' tendency to identify the periodicity of the stimulus by attending to the dominance region, and to perceive pitch from the representation of stimulus frequency within that region. Independent evidence from studies of nonspeech tones and vowels supports the general conclusion that frequency in the dominance region causes apparent pitch, even for natural speech. Hence, the explanation of sinusoidal intonation that we offer is that these atypical stimuli are evaluated perceptually in essentially the same manner as nonspeech tonal complexes and speech sounds are.

However, the fact that the subjects chose Tone 1 consistently as the best match to apparent pitch does not mean that Tone 1 was causing the pitch percept. To support this characterization of the perception of sinusoidal intonation, we would have to determine that subjects do not select Tone 1 when it is absent from the tonal sentence. If subjects selected Tone 1 as the match to intonation *only* when it was present in the sentence, then we would have reasonable grounds to support our stimulus-based hypothesis of the phenomenon. Otherwise, if subjects continued to prefer Tone 1 to other candidate tones when that tone was omitted from the sinusoidal sentence, we would necessarily conclude that intonation occurred through a form of auditory induction, however similar this induced pattern was to the pitch contour of Tone 1. Experiment 3 evaluated this possibility by presenting a test of differential similarity in which the sentences to be matched contained either the three tones corresponding to the first three formants or merely the tones corresponding to the second and third formants, omitting the first.

## Method

**Subjects.** Twenty-one listeners were selected from the student population of Barnard and Columbia colleges, again. None had participated previously in experiments of this nature. They were paid for their participation.

**Stimuli.** The three-tone sinusoidal replicas of the sentence "I read a book today" prepared for Experiment 2 provided the basis for all stimuli in this test. Three versions of the sentence were used. The first was the uptilted amplitude replica, in which the tone amplitudes were the inverse of the natural case of formants. Tone 1 had the least power, and Tone 3 the most. The second sentence was the pattern consisting only of Tones 2 and 3 of this replica. This two-tone pattern was equated informally, by the authors, for loudness equal to the three-tone pattern. Note that Tone 1 is omitted from this pattern. The third sentence was the three-tone replica, preserving the natural amplitude relations among the tones but presented at low power, again to serve as a check on the outcome. The three single-tone patterns from Experiment 2 were used as alternative pitch contours in this test of differential similarity.

**Procedure.** Listeners were instructed to identify the sentence melody of the sinusoidal sentence presented first on each trial, and then to select the better match to that sentence melody from the two lagging single-tone alternatives. The subjects were urged not to omit judgments. The choices were scored in pencil in specially prepared response booklets.

There were three different combinations of alternatives, counterbalanced for order of presentation: Tone 1 vs. Tone 2, Tone 1 vs. Tone 3, and Tone 2 vs. Tone 3. Each trial type was presented 20 times in random order with each of the two sentence versions, the three-tone one and Tones 2 and 3. A third sentence, normal-quiet, occurred 12 times in this test paired only with Tone 1 vs. Tone 2 alternatives. The test, then, consisted of 132 trials. Within a trial, the three patterns were separated by intervals of 1 sec. Trials were separated by 3 sec, with 8 sec between blocks of 12 trials.

## Results and Discussion

The results of the similarity judgments are shown in Figure 5. It is clear that subjects once again selected Tone 1 when it occurred as a component of the sentence. In the case of the sentence containing only Tones 2 and 3, though, subjects instead preferred Tone 2 to Tone 1 as the best match for the sentence intonation. This outcome corresponds to a highly significant interaction term in the analysis of variance [$F(2,40) = 52.4$, $p < .001$]. The subjects also preferred Tone 2 when it was pitted against Tone 3 in the context of the two-tone sentences. Overall, subjects reported that sentence pitch was matched best by Tone 1 only when that tone was a component of the sentence.

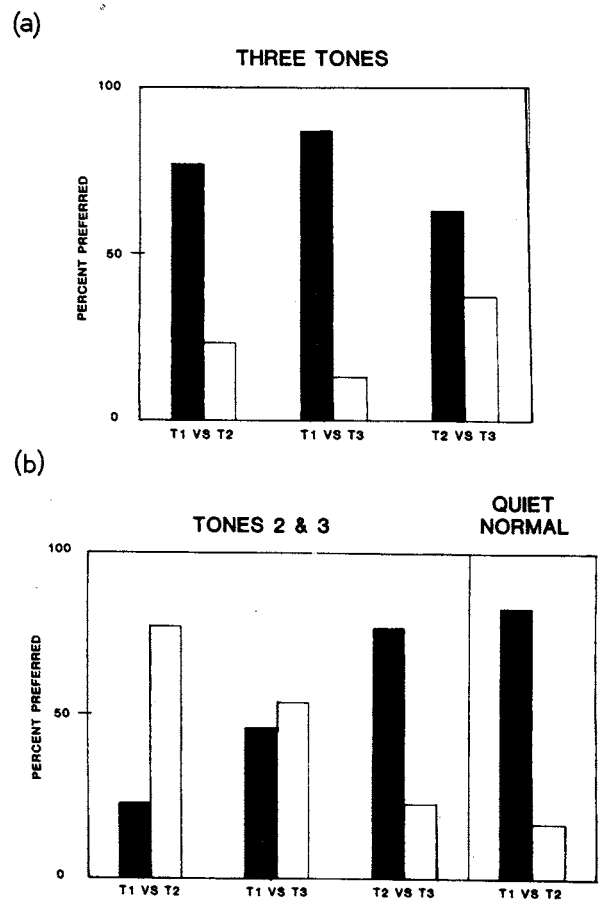


**Figure 5. Results of Experiment 3. (a) Differential similarity data for the three-tone sentence with flat rolloff. (b) Differential similarity data for the two-tone and quiet-normal conditions.**

This third experiment is encouraging with respect to the hypothesis we offered about sinusoidal intonation. The subjects' treatment of these anomalous signals appears to be similar to their treatment of speech signals. It is as if the segmental information was obtained from the formant-like frequency variation of the tones and intonational information was provided by the periodicity within the dominance region. This occurred despite the congruence of these two kinds of information in the pattern of frequency variation of Tone 1.

However, to establish the appropriateness of this application of the dominance region notion, we must perform one final test. This is necessitated by the kind of evidence we have obtained so far on the predominance of Tone 1 in producing the apparent intonation. Although our experiments have shown that listeners consistently judge this tonal component to be most like the sentence melody of a sinusoidal utterance, we have not separated two aspects of this tone within the three-tone pattern that composes a sentence. In the three tests that we report, the tone corresponding to the first formant has been both the tone within the dominance region and the tone with the lowest frequency, overall, in the three-tone complex. Because of this fact, we cannot distinguish empirically between the dominance region hypothesis and a lowest frequency component hypothesis. To do so requires a test in which the subjects evaluate a sentence that contains tonal components falling in the dominance region, below the dominance region (with frequencies $< 400$ Hz), and above the dominance region (with frequencies $> 1000$ Hz). We can predict the outcome based on Experiments 1-3: When subjects listen to such a sentence, they should either attend to the tone within the critical frequency range for perceiving intonation, which would encourage the dominance region explanation that we have proposed, or they should prefer the lowest frequency tone, which would falsify the dominance region hypothesis, although in a manner consistent with the findings we have noted throughout this investigation. This test is the topic of Experiment 4.

## EXPERIMENT 4

The original rationale for the dominance region was that the auditory system gets the stimulus for pitch where the harmonics are resolved the best. At this juncture, we have shown the superiority of Tone 1 (corresponding to the first formant) compared with simultaneously occurring tones with higher frequencies. Additionally, the dominance-region hypothesis predicts that listeners should also reject tones falling below the dominance region. To perform this test of the claim, we returned to the natural utterance of our familiar test sentence, and analyzed its fundamental frequency pattern. From this analysis, a new set of sine-wave synthesis parameters was created to form a tone with a pattern of frequency variation matching the natural fundamental frequency contour. These values were used in combination with the three-tone replica to generate a four-tone sentence, comprising a "fundamental frequency" tone and the three "formant" tones, as well as the additional single-tone alternative to use in the similarity test format.

In the four-tone sentence that subjects evaluated, the tone matching the fundamental frequency contour falls below the dominance region. If the likeness of the first formant tone to the apparent sinusoidal intonation is based on its occurrence within the critical frequency range, then we may expect listeners to reject the fundamental frequency tone no less consistently than they have rejected the second and third formant tones in Experiments 1, 2, and 3. In other words, a tone representing a fundamental frequency pattern from a natural utterance should ironically not provide information for sentence melody in this case, despite the naturalness of its pattern of variation and the appropriateness of its occurrence in the normal frequency range of the fundamental frequency.

### Method

**Subjects.** Twenty-four listeners participated in this study. They each reported a normal history of speech and hearing function, and had not previously been introduced to synthetic speech or sine-wave materials. Our subjects were student volunteers who received course credit in exchange for taking this brief test.

**Stimuli.** The sentence presented to subjects in this test was composed of four tones: Tone 0, corresponding to the fundamental frequency (commonly termed F0) and overall amplitude of the original natural utterance of "I read a book today," on which we patterned the sine-wave sentences reported in the prior two experiments; and Tone 1, Tone 2, and Tone 3, each corresponding to the pattern of center-frequency and amplitude variation of the first three formants. The values of the fundamental of the natural utterance were obtained by employing the cepstral method of pitch extraction on the sampled data, and were converted to sine-wave synthesis parameters by including amplitude values varying in imitation of the overall energy of the natural utterance. The pattern of frequency variation of Tone 0 is shown in Figure 6a. The four-tone pattern formed by combining Tone 0 with the three tones that replicate formant variation preserved the natural spectral amplitude rolloff, as shown in Figure 6b.

The test stimuli also included the four sinusoidal components realized as single-tone patterns, to be used as alternative pitch stimuli in the similarity test. Each of the tones was resynthesized in isolation, and the four were equated for loudness.

As before, the test sequence was recorded on audiotape and presented to listeners via playback and calibrated headsets. An average listening level of 72 dB SPL was used.

**Procedure.** A test of apparent similarity was used again in this experiment. Each trial consisted of three sinusoidal patterns: first, the four-tone sentence pattern, followed by two single-tone patterns. There were six different trial types, exhausting the possible comparisons among the four single-tone candidates: Tone 0 vs. Tone 1, Tone 0 vs. Tone 2, Tone 0 vs. Tone 3, Tone 1 vs. Tone 2, Tone 1 vs. Tone 3, and Tone 2 vs. Tone 3. Each was presented in two orders to counterbalance the occurrence of alternatives. Altogether, the test consisted of the six trial types presented 14 times each, including counterbalancing, composing a sequence of 84 trials.

On each trial, the subjects were instructed to identify the sentence melody of the first sinusoidal pattern, and then to select the

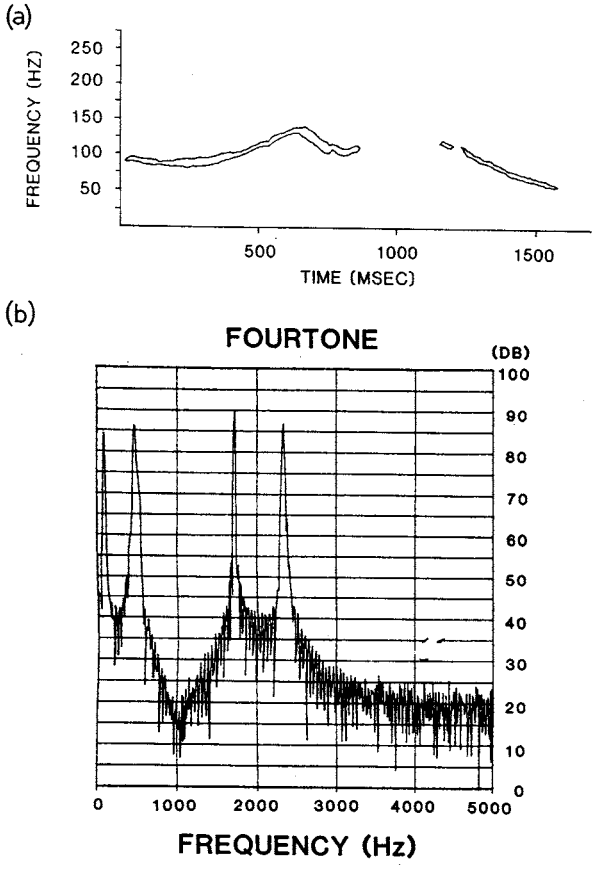


**Figure 6. (a) The frequency pattern of Tone 0, which reproduced the fundamental frequency pattern of the natural utterance "I read a book today" in sinusoidal form. (b) A representative section through the four-tone sentence pattern.**

better match of the two lagging alternative patterns. Omissions were discouraged. The judgments were reported with pencil and paper using specially prepared booklets.

### Results and Discussion

The histograms in Figure 7 describe the results of the similarity test. Tone 1, corresponding to the first formant, was once again preferred to every other candidate tone. Tone 2 was judged more similar to the intonation pattern than was Tone 3, an unanticipated effect. And, most critically, the subjects rejected Tone 0 consistently when it was an alternative paired with Tone 1, indicating that the impression of sentence melody was stable. These results were confirmed in the analysis of variance of similarity scores [$F(5,115) = 40.1$, $p < .001$] and by Scheffé post hoc means tests.

The pattern of results of Experiment 4 clearly confirms the appropriateness of the dominance region hypothesis for the phenomenon of sinusoidal sentence intonation. In fact, the congruence of segmental and intonational information in the sinusoidal case of Tone 1 permits us to support a proposal about auditory analysis of natural speech: Fundamental periodicity is represented in the auditory system based on harmonics detected within the dominance region and not on attention to the fundamental itself. Because Tone 1 occurs within the range of this normal region for detecting periodicity in the waveform, it seems to be treated as *the* principal stimulus for pitch perception.

## GENERAL DISCUSSION

Prosody is a perceptual dimension of utterances that is not caused by variation in any single physical dimension of the acoustic signal. The listener is likely to treat the duration, amplitude, and fundamental period of portions of the speech signal as changes in the rhythm, meter, and organization of the linguistic utterances that perception defines. One aspect of prosody is intonation, or sentence melody. The problem for the theorist is to identify the relations among the quite dissimilar physical ingredients that produce impressions of intonation in some cases, but create impressions of duration, or loudness, or lexical stress, or perhaps syntactic constituent boundaries, in others. In addition to the effects of these physical variables, perception of intonation has been viewed

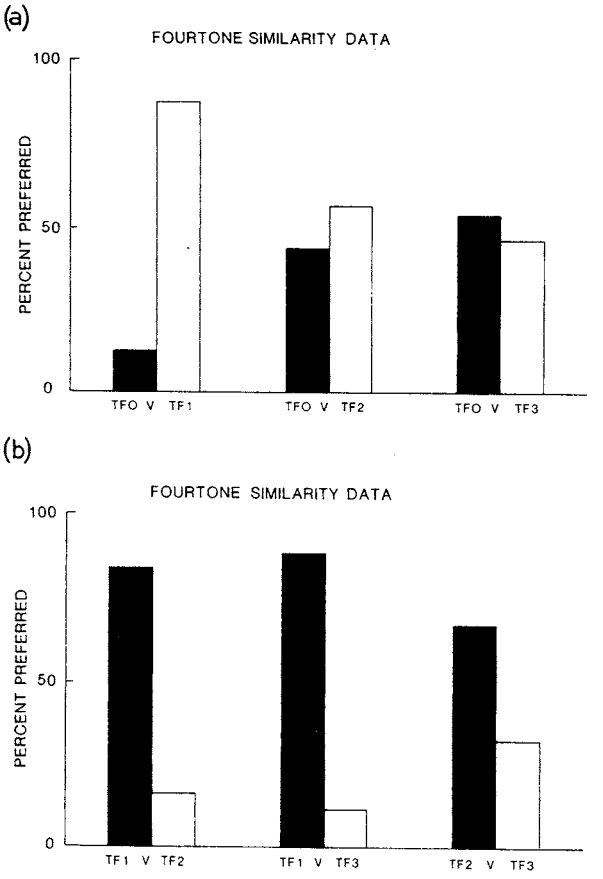


**Figure 7. Results of Experiment 4. (a) Differential similarity data for comparisons involving Tone 0. (b) Differential similarity data for remaining comparisons.**

as a process that refers to linguistic knowledge, because judgments of intonation often reflect lexical properties (Lieberman, 1965; but see Lea, 1979).

Given the intricate interplay of physical and perceptual components in prosodic perception, it seems anticlimactic to assert that the perception of intonation is based principally on fundamental frequency, in some instances necessarily so (Abramson, 1972). However, intonation is potentially determined from integrated energy or from frequency variation in the third and fifth formants in whispered sentences (Meyer-Eppler, 1957), which lack contours of fundamental frequency. As such, the whispered utterance is the most reasonable precedent for sinusoidal sentence perception. A sinusoidal replica also lacks a fundamental frequency of excitation common to its tonal components, and therefore we might have expected it to be treated in a manner similar to that of a whispered sentence. Instead, we found consistent perceptual reliance on the portion of the signal within the dominance region as the primary ingredient to intonation, much as occurs for normal utterances.

We cannot yet define a principle by which intonation is variously derived from the fundamental, or the amplitude envelope, or the higher formant frequency changes. Because our exploratory studies probed this phenomenon at the sentence level, neither have we determined the extents of the likely influence of duration, amplitude, and relative frequency change, on the one hand, or of lexical access, constituent structure, and the encoding of intonation in memory, on the other. Each of these factors may be suspected of moderating the effect of fundamental frequency. Even if these other influences are slight, we may nevertheless expect intonation to differ from the fundamental frequency pattern (Hadding-Koch & Studdert-Kennedy, 1964). With these cautions in mind, we propose that our investigation describes the perceptual registration of the strongest influence on intonation, the fundamental frequency.

The studies reported here lead us to conclude that speech signals are analyzed for fundamental frequency in the dominance region, coincidentally, the region of the first formant, as Greenberg (1980) hypothesized on the basis of studies of the strength of periodicity in auditory evoked potentials with synthetic vowels. It is somewhat ironic that sinusoidal signals, clearly unnatural in vocal timbre, provide evidence on this question. But, if the auditory system ordinarily detects periodicity from the harmonics in the dominance region, then when it fails to find harmonics it seems nevertheless to represent the pitch of a complex signal by its period in this region. A sinusoidal sentence is a kind of exceptional stimulus that tests the rule, and confirms it.

Is the intonation of sinusoidal sentences the result of periodic acoustic structure subsequently transformed by duration and loudness (or by segmental and morphological structure)? If sinusoidal signals and natural speech are analyzed in a common manner, as we claim, then we may certainly expect sinusoidal intonation to be affected by acoustic and linguistic properties besides frequency of the tone in the critical range. For the present, though, the evidence suggests that the primary correlate of sinusoidal intonation is the tone that reproduces the frequency variation of the first formant. And, while this outcome is revealing about the perception of natural speech, it also supports the contention that sinusoidal replicas of utterances are perceived like ordinary phonetic signals.

## REFERENCES


Abramson, A. S. (1972). Tonal experiments with whispered Thai. In A. Valdman (Ed.), *Papers in linguistics and phonetics in memory of Pierre Delattre* (pp. 31-44). The Hague: Houton.

Abramson, A. S., & Lisker, L. (1965). Voice onset time in stop consonants: Acoustic analysis and synthesis. In D. E. Commins (Ed.), *Proceedings of the 5th International Congress of Acoustics* (A-51). Liege: G. Thone.

Bailey, P. J., Summerfield, A. Q., & Dorman, M. (1977). On the identification of sine-wave analogues of certain speech sounds. *Haskins Laboratories Status Report on Speech Research*, **SR-51/52**, 1-25.

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

Catford, J. C. (1977). *Fundamental problems in phonetics*. Bloomington: Indiana University Press.

Collier, R., & t'Hart, J. (1975). The role of intonation in speech perception. In A. Cohen & S. Nooteboom (Eds.), *Structure and process in speech perception*. New York: Springer.

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

Fant, C. G. M. (1956). On the predictability of formant levels and spectrum envelopes from formant frequencies. In M. Halle, H. Lunt, & H. MacLean (Eds.), *For Roman Jakobson* (pp. 109-120). The Hague: Mouton.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, **1**, 120-152.

Greenberg, S. (1980). *Temporal neural coding of pitch and vowel quality* (Working Papers in Phonetics, no. 52, pp. 1-183). Los Angeles: U.C.L.A.

Grunke, M. E., & Pisoni, D. B. (1982). Perceptual learning of mirror-image acoustic patterns. *Perception & Psychophysics*, **31**, 210-218.

Hadding-Koch, K., & Studdert-Kennedy, M. (1964) An experimental study of some intonation contours. *Phonetica*, **11**, 175-185.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, **25**, 105-113.

Joos, M. A. (1948). Acoustic phonetics. *Language*, **24**(Suppl., Language Monographs 23), 1-136.

Lea, W. A. (1979). Testing linguistic stress rules with listeners' perceptions. In J. J. Wolf & D. H. Klatt (Eds.), *Speech communication papers* (pp. 415-418). New York: Acoustical Society of America.

Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa*, **7**, 107-122.

Lehiste, I., & Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America*, **31**, 428-435.

LEHISTE, I., & PETERSON, G. E. (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America*, **33**, 419-423.

LICKLIDER, J. C. R. (1956). Auditory frequency analysis. In C. Cherry (Ed.), *Information theory*. New York: Academic Press.

LIEBERMAN, P. (1965). On the acoustic basis of the perception of intonation by linguists. *Word*, **20**, 40-54.

LIEBERMAN, P. (1967). *Intonation, perception, and language*. Cambridge: M.I.T. Press.

MEYER-EPPLER, W. (1957). Realization of prosodic features in whispered speech. *Journal of the Acoustical Society of America*, **29**, 104-106.

MORTON, J., & JASSEM, W. (1965). Acoustic correlates of stress. *Language and Speech*, **8**, 159-181.

PLOMP, R. (1967). Pitch of complex tones. *Journal of the Acoustical Society of America*, **41**, 1526-1533.

REMEZ, R. E., RUBIN, P. E., & CARRELL, T. D. (1981). Phonetic perception of sinusoidal signals: Effects of amplitude variation. *Journal of the Acoustical Society of America*, **69**, S114.

REMEZ, R. E., RUBIN, P. E., PISONI, D. B., & CARRELL, T. D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-950.

RITSMA, R. J. (1967). Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America*, **42**, 191-198.

SCHOUTEN, J. F. (1940). The residue and the mechanism of hearing. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, **43**, 991-999.

SCHWAB, E. C. (1981). *Auditory and phonetic processing for tone analogs of speech*. Unpublished doctoral dissertation, State University of New York at Buffalo.

STREETER, L. A. (1978). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, **64**, 1582-1592.

SUMMERFIELD, Q., & HAGGARD, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial position. *Journal of the Acoustical Society of America*, **62**, 435-448.

## NOTES

1. A pure tone is not a formant. A sinusoid is defined by the function $y = a \cdot \sin x$, and may occur at any frequency within the audible range. A formant is a natural resonance of the vocal tract, and its frequency is defined as the peak of the spectrum envelope drawn to enclose the harmonics produced by the excitation of the vocal tract (Fant, 1956). Although we have constructed sinusoids that imitate the pattern of formant center-frequency variation, they do not also imitate the acoustic structure of formants, by this definition. For a basic discussion of the physical acoustics of speech, see Joos (1948).

2. The intonation of a sentence is its pitch contour (Catford, 1977), although this definition is perceptually troublesome. This is so because the term *pitch* is traditionally used to refer to that perceptual impression correlated with fundamental frequency. Intonation is also correlated mainly with fundamental frequency, although pitch applies to speech and nonspeech and intonation more narrowly applies to speech exclusively. In view of this, is sentence intonation the product or the equivalent of sentence pitch contour? The fact that aspects of signal duration and power intrude on the perception of both intonation and pitch argues that both terms name the same attribute. The influence of lexical structure in judging sentence melody argues against any simple equivalence, although it by no means warrants that a separate auditory impression of pitch contributes to the impression of intonation. [Linguists have occasionally combined the analysis of intonation and word stress (reviewed by Lieberman, 1967), although to do so does not dismiss the phenomenon of sentence pitch—it simply adds another problem to consider.] Our present use of the term, then, refers to the fact that sentence "pitch contour," sentence "melody," and sentence "intonation" seem to indicate the same aspect of spoken sentences, although its perceptual derivation is difficult to resolve.

(Manuscript received July 21, 1983;
revision accepted for publication March 7, 1984.)