Journal of Experimental Psychology:
Human Perception and Performance
1987, Vol. 13, No. 1, 40-61

Perceptual Normalization of Vowels Produced by Sinusoidal Voices

Robert E. Remez

Philip E. Rubin

Lynne C. Nygaard

William A. Howell

0096-1523/87/$00.75

Journal of Experimental Psychology:
Human Perception and Performance
1987, Vol. 13, No. 1, 40-61

0096-1523/87/$00.75

Perceptual Normalization of Vowels Produced by Sinusoidal Voices

Robert E. Remez
Barnard College

Philip E. Rubin
Haskins Laboratories, New Haven, Connecticut

Lynne C. Nygaard
Barnard College

William A. Howell
Columbia College

When listeners hear a sinusoidal replica of a sentence, they perceive linguistic properties despite the absence of short-time acoustic components typical of vocal signals. Is this accomplished by a postperceptual strategy that accommodates the anomalous acoustic pattern ad hoc, or is a sinusoidal sentence understood by the ordinary means of speech perception? If listeners treat sinusoidal signals as speech signals however unlike speech they may be, then perception should exhibit the commonplace sensitivity to the dimensions of the originating vocal tract. The present study, employing sinusoidal signals, raised this issue by testing the identification of target /bVt/, or b-vowel-t, syllables occurring in sentences that differed in the range of frequency variation of their component tones. Vowel quality of target syllables was influenced by this acoustic correlate of vocal-tract scale, implying that the perception of these nonvocal signals includes a process of vocal-tract normalization. Converging evidence suggests that the perception of sinusoidal vowels depends on the relations among component tones and not on the phonetic likeness of each tone in isolation. The findings support the general claim that sinusoidal replicas of natural speech signals are perceptible phonetically because they preserve time-varying information present in natural signals.

Does speech perception depend on the occurrence of specific acoustic elements within the signal spectrum? The customary conceptualizations of phonetic perception have assumed so and have sought to describe the manner in which elementary acoustic cues bring about the perception of phoneme sequences. For example, isolatable spectral elements, such as brief formant frequency transitions, momentary aperiodicities, and low frequency murmurs, are said to be correlates of the perception of consonantal attributes. These particular three are responsible for the perception of articulatory place, voicing, and nasal manner, respectively (Delattre, Liberman, & Cooper, 1955; Zue & Schwartz, 1980). Other descriptions of the transformation of acoustic structure to phonetic properties have relied on momentary spectral shapes (Stevens & Blumstein, 1981) or spectral sequences (Klatt, 1979) or cue-weighting techniques (Massaro, 1972) or network arrangements (Eimas & Miller, 1978; Elman & McClelland, 1985). Though these mechanisms have been entertained as rivals, among others, for explaining speech perception, they all begin in the same way: They identify or filter particulate elements of the acoustic–auditory pattern, assuming them to be the ingredients essential for the perceptual process.

A few proposals, including one of our own, have departed from this point of view (Bailey & Summerfield, 1980; Jenkins, Strange, & Edman, 1983; Liberman, 1970; Remez, Rubin, Pisoni, & Carrell, 1981). In our work, we have employed tonal analogs of speech composed of three or four time-varying sinusoids, each one reproducing the frequency and amplitude variation of a vocal resonance in a natural utterance. A sinusoidal sentence pattern retains the overall configuration of spectral variation of the natural utterance on which it is modeled. But it is devoid of aperiodicities, fine-grained regular (glottal) pulses, harmonic series, and broadband formant structure, which compose the natural speech signal, and therefore lacks the rich assortment of acoustic particles that typically receive the emphasis in explanations of speech perception. Perceptual tests employing sinusoidal sentences offer an opportunity to test the effects of structured acoustic patterns, independent of the assortment of acoustic elements found in speech signals.

Listeners judge sinusoidal signals to be unlike speech, as might be suspected from considering the short-time acoustic properties of tonal signals. Acoustically and perceptually, sinusoidal signals are grossly unnatural, and naive test subjects who are told only to identify "sounds" tend to perceive sinusoidal sentences merely as several covarying tones (Remez et al., 1981). This outcome is predicted by conventional explanations of speech perception based on acoustic elements. However, when instructed to listen for a linguistic message in the tonal patterns, such test subjects often succeed, though only when the tonal configuration abstractly represents a complex resonance pattern; a single tone reproducing the second formant, for example, is perfectly untranscribable.

Despite the fact that many listeners have little difficulty in comprehending sinusoidal patterns that replicate several formants, as many as a third of the listeners may fail utterly to do so. They report instead that the tone-complexes do not cohere,

For sponsorship, the authors thank the National Institute of Neurological and Communicative Disorders and Stroke (Grant NS-22096 to Robert E. Remez) and the National Institute of Child Health and Human Development (Grant HD-01994 to Haskins Laboratories). For help and criticism, the authors thank Randy Diehl, Louis Goldstein, Jan Rabinowitz, Jim Sawusch, and Michael Studdert-Kennedy.

Correspondence concerning this article should be addressed to Robert E. Remez, Department of Psychology, Barnard College, 3009 Broadway, New York, New York 10027.

which perhaps encourages the view that different and mutually exclusive perceptual organizations are possible for sinusoidal signals (Bailey, Summerfield, & Dorman, 1977; Best, Morrongiello, & Robson, 1981; Cutting, 1974; Remez et al., 1981; Remez, in press; Williams, Verbrugge, & Studdert-Kennedy, 1983). However, we have yet to identify a common independent factor differentiating the majority of listeners who are able to transcribe the patterns from the minority who are unable.

Our general hypothesis is that sinusoidal replicas of speech convey linguistic information in patterned variation, for few of the acoustic details typical of natural speech remain to provide the basis for phonetic perception. If our interpretation is plausible, then these phenomena are evidence of a kind of acoustic information in speech perception that is time-varying and relatively independent of elemental constituents. The sufficiency of time-varying information presents a strong challenge to perceptual accounts that describe the information in acoustic signals as discrete cues to be extracted from the speech stream like raisins from pudding. Moreover, phonetic perception of tone complexes resists description by prototype models that favor schematizations of typical acoustic elements or their auditory representations (for example, Massaro, 1972; Samuel, 1982).

Time-Varying Phonetic Information

The patterns of variation in sinusoidal sentences are derived from natural utterances, which suggests a clue to understanding the skill shown by listeners transcribing these spectrally unfamiliar tonal patterns. Sinusoidal signals may preserve phonetic information present in the natural signals on which they are modeled. Perhaps such information is available ordinarily in the coherent variation of natural signals. Certainly, listeners require no special training to perform adequately in the transcription test, which implies that sinusoidal signals replicate the information available from coherent signal variation despite the gross short-time differences between sinusoidal and natural spectra. Sinusoidally presented linguistic attributes would therefore be perceptible because this time-varying information is treated in an ordinary way—as occurs with time-varying phonetic information in a natural signal.

As attractive as that conclusion is for us, there is an alternative hypothesis that may fare as well in accounting for the findings. Suppose that sinusoidal imitations of speech signals merely preserve aspects of acoustic structure that are irrelevant to phonetic identification. In the absence of the natural acoustic products of vocalization, the listener may exploit the residual, rough resemblance of sinewave variation to speech signals in order to imagine a linguistic sequence that plausibly fits the tones. If so, then we may expect transcription of sinusoidal replicas of sentences to occur via postperceptual compensation for the failure of phonetic perception, subjecting the nonspeech percept to inference, analogic reasoning, and outright guesswork in concocting a linguistic likeness.

An example from the archives that appears to fit this postperceptual description is contributed by Newton, who wrote:

> Soe ye greatest cavity in ye mouth being first made in ye throate & thence by degrees moved towards ye lipps further from ye larinx causes ye pronunciation of ye vowells in order y i e a o ω u w. The filling of a very deep flaggon with a constant streame of beere or water sounds ye vowells in this order w u ω o a e i y. (Newton, 1665, quoted by Ladefoged, 1967, p. 65)

It seems that Newton heard the changing pitch of the flagon's resonance as a series of vowels—[u] when the flagon was empty and had low-pitched resonance, through [i] when the flagon was almost full and had high pitch. Surely, Newton never thought the flagon spoke vowels to him, yet his transcription performance would presumably have been reliable. This is a credible instance, then, in which an auditory experience is likened to speech by deliberate afterthought and is distinct from the immediate perception of speech.

Taking Newton's observation into account, our immediate question may be clearly posed: Is perception of sinusoidal signals like speech perception or rather like listening to nonphonetic sounds and then inventing a sentence to match a nonspeech pattern? The answer determines whether research employing tonal analogs of natural signals contributes a valid approach to studying the perception of speech.

Unfortunately, there is no direct way to establish whether perception of tone complexes is primarily phonetic or nonphonetic. We cannot appeal to a definitive test to discover whether acoustic patterns are perceived to be composed of speech sounds or perceived to be nonphonetic but less unlike some speech sounds than others. We may obtain an indirect answer by observing whether phonetic identification of a sinusoidal signal is affected by the implied dimensions of the vocal tract, which seems to produce the tonal sentence. Listeners should demonstrate perceptual normalization of vocal-tract dimensions only if there is information in the sinusoidal pattern to mark the tone patterns as implicitly vocal and only if the tone variations possess sufficient structure to evoke the low-level perceptual evaluation of the sinusoidal voice relative to the range of potential talkers (see Fant, 1962; Joos, 1948; Nearey, 1978). In other words, the solution to the puzzle can be obtained by seeing whether phonetic transcription of sinusoidal signals exhibits perceptual normalization of the acoustic properties correlated with intertalker variation in vocal-tract dimensions. If so, this would be evidence for a basic and early perceptual function keyed to speech sounds, from which we could infer that sinusoidal transcription is an instance of speech perception.

Experiment 1

Normalization of Vocal-Tract Dimensions

Our test makes use of a classic finding by Ladefoged and Broadbent (1957), that the perception of speech entails a process which accommodates the acoustic signal properties attributable solely to variation among individual talkers, independent of linguistic factors. The root of this acoustic variation is the corresponding variation in vocal anatomy and in the control of the articulators (Bricker & Pruzansky, 1976; Joos, 1948). Briefly, when different talkers speak the same word, the specific frequencies traversed by the formants differ from one talker to another. Some vocal tracts are longer, some shorter, producing a kind of scalar variation in formant patterns observed across a range of talkers employing the same linguistic elements (see also Fant, 1966; Peterson & Barney, 1952). Differences in vocal gestures and corresponding formant trajectories may also result

from differences in articulatory control (Ladefoged, 1980). The consequence is plain; the perceiver cannot identify vowels simply from momentary formant frequency values—vowels must first be implicitly rescaled with reference to the talker's formant range, in other words, with respect to the capability of the vocal tract that produced them. Because the formant pattern of one talker's *bet* can be identical to that of another's *but,* this rescaling ensures that the word produced by the talker is perceived as such, all other things being equal.

Ladefoged and Broadbent (1957) made perceptual normalization an empirical issue using the technique of speech synthesis. Initially, they produced synthetic imitations of a natural utterance, "Please say what this word is," along with approximations of the words *bit, bet, bat,* and *but.* The synthetic sentence reflected the vocal-tract dimensions of the talker who produced the natural utterance from which the sentence synthesis was derived. However, transpositions of individual formant tracks of the sentence pattern in the frequency domain gave it characteristics of vocal tracts of rather different dimensions. Ladefoged and Broadbent lowered (by 25%) or raised (by 30%) the first or second formants (or both) to create impressions of different talkers from the original sentence pattern. In the normalization test, listeners labeled the target syllables that differed only in the identity of the vowel. Each of the four targets was presented with the untransposed sentence frame, to establish baseline labeling, and was also presented with one or more transposed sentences as the frame. In the latter cases the distributions of judgments often differed from those that had been observed with the untransposed, natural sentence pattern as the frame.

The experimental outcomes were rationalized by assuming that perceptual rescaling occurred on the basis of the leading sentence frame and was revealed in the effects of different frames on identification of the same lagging target syllable. The perceiver presumably identified the target vowel by reference to the range of formant variation encountered in the frame. This seemed reasonable because the capability possessed by any talker for producing formant frequency excursions is limited by the size of the vocal cavities, and the formant values associated with any specific vowel posture or gesture are governed therefore by the overall range.

The contingency of vowel identity on the scale of the originating vocal tract has been described as adaptation to the personal dimension of the information in speech signals (Ladefoged & Broadbent, 1957); as perceptual normalization for the talker (Gerstman, 1968; Nearey, 1978) based on inferences about the physical characteristics of the sound source (Fourcin, 1968); as calibration or mapping of the talker's vowel space (Dechovitz, 1977a; Verbrugge, Strange, Shankweiler, & Edman, 1976); as vowel normalization (Disner, 1980; Sussman, 1986; Wakita, 1977); and as vocal-tract normalization (Joos, 1948; Rand, 1971; Shankweiler, Strange, & Verbrugge, 1977; Summerfield & Haggard, 1973). Susceptibility to vocal-tract normalization may serve as an index of the phonetic perceptual effects of sinusoidal signals, because it is unlikely that effects attributable to normalization arise postperceptually. First, normalization in speech perception is rudimentary and is evident early in development (Kuhl, 1979; Lieberman, 1984). Second, though normalization may be automatic and effortless in speech perception, it is implausible to suppose that listeners are equally adept in deliberate guessing, postperceptually, about articulatory geometry. The history of several centuries of such intuitive efforts antedating the use of modern techniques shows that the dimensions of speech production were commonly described imprecisely or misconstrued (Ladefoged, 1967; chapter 2). Third, it is doubtful that the detailed precategorical acoustic structure presumably required for a postperceptual approximation of normalization survives long enough in memory to be useful in this regard (for example, Darwin & Baddeley, 1974). If listeners show evidence of normalizing sinewave replicas, then the perceptual explanation ordinarily offered for this kind of effect with synthetic speech would therefore apply with equal force to the sinusoidal case.

Therefore, the first experiment that we report here attempted to obtain a verdict about sinusoidal replication: Do subjects perform the transcription of sinusoidal signals by relying on phonetic perception or on a postperceptual elaboration of auditory perception? If we observe effects of normalization, then we may say that perception of tonal analogs of speech occurs in a manner akin to ordinary phonetic perception.

Method

Acoustic Test Materials

Six versions of the sentence, "Please say what this word is," and the four target words *bit, bet, bat,* and *but* were prepared by the technique of sinewave replication of natural speech signals. An adult male speaker (RER) of Northeast American English produced a single utterance of the sentence and of each of the four targets in a sound attenuating chamber (I.A.C.). These were recorded with a condenser microphone (Shure SM-78) on audiotape (Scotch #208), using a half-track recorder (Otari MX5050).

Spectra were determined by digitizing the utterances and analyzing the records. After filtering the playback signal by low-passing at 4.5 kHz, it was sampled at 10 kHz with 12-bit resolution and stored on a DEC VAX 11/780-based system. The values of the frequencies and amplitudes of the formants were computed at intervals of 10 ms throughout each natural utterance, using the technique of linear prediction (Markel & Gray, 1976). In turn, these values were used to control a sinewave synthesizer (Rubin, 1980), which computes waveforms of signals generated by multiple independent audio-frequency oscillators. Four sinusoids were used to replicate the sentence pattern, one for each of the three lowest oral formants and a fourth sinusoid for the fricative resonance when appropriate. Spectrographic representations and spectral sections are shown in Figures 1 and 2 for the natural sentence that served as the model, and in Figures 3 and 4 for the sinusoidal replica. In the present test this sentence, which replicates the linear prediction estimates of the resonant frequencies of the natural utterance, functions as the untransposed *natural* precursor. Three sinusoids were used to replicate each of the [bVt] targets. Table 1 contains the frequency values for the three tones of each target word determined at the syllable nuclei.

Five additional versions of the sentence were prepared to parallel the conditions used by Ladefoged and Broadbent by transposing the frequency values of the sinusoid that followed the first or the second formant of the natural utterance. They are (a) with Tone 1 lowered by 25%; (b) with Tone 1 raised by 30%; (c) with Tone 2 lowered by 25%; (d) with Tone 2 raised by 30%; and, (e) with Tone 1 lowered and Tone 2 raised together, respectively, by 25% and 30%.

Test sequences were prepared on the VAX and recorded on half-track audiotape (Scotch #208) following digital-to-analog conversion. The acoustic materials were delivered binaurally to test subjects via playback

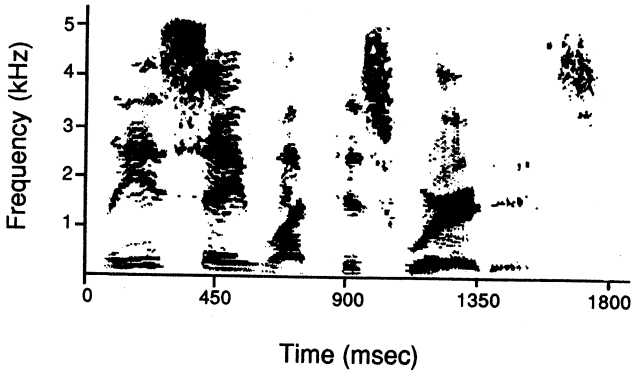

Figure 1. Spectrographic representation of the natural sentence, "Please say what this word is."

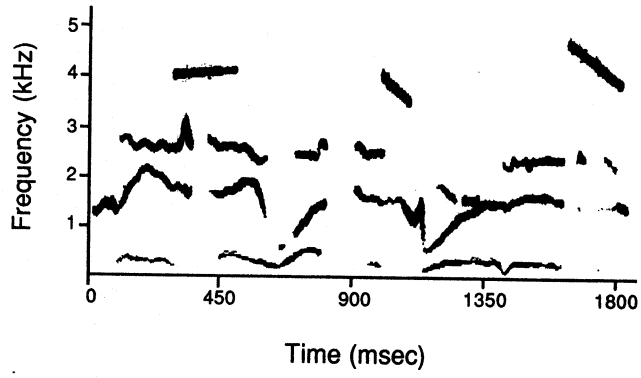

Figure 3. Spectrographic representation of the sinewave pattern derived from the natural sentence "Please say what this word is," in which natural formant frequencies and amplitude values were used to control digital audiofrequency oscillators. (Three tones follow the oral formants; a fourth tone replaces the fricative formant when appropriate.)

tape recorder (Otari MX5050), power amplifier (Crown D-75), and matched headsets (Telephonics TDH-39, 300 ohm/channel), attenuated approximately to 60db (SPL).

Procedure

A testing session included four parts. The first was a warm-up in which eight sinusoidal sentences of varying phonetic composition were presented for transcription. This segment was used simply to accustom the listeners to the unusual timbre of the signals and was not scored.

The second segment was the normalization test replicating the procedure of Ladefoged and Broadbent, in which target syllables were presented for identification within sentence frames. This test consisted of 11 trials with the same format: a sentence frame (2,400 ms), a silent interval (500 ms), and a target syllable (on the average, 140 ms). Successive trials were separated by 10 s of silence. There were 11 different conditions in this test, one trial per condition, each trial separated from the preceding and following trial by a long silence. This procedure was adopted from Ladefoged and Broadbent and aimed to prevent subjects from developing familiarity with the small set of syllables and identifying them by rote rather than by perceiving the vowels. The target *bit* was presented with two frames—natural and Tone 1 lowered; *bet* was presented with four frames—natural, Tone 1 lowered, Tone 1 raised, and Tone 1 lowered and Tone 2 raised; *bat* was presented with three frames—natural, Tone 1 raised, and Tone 2 raised; and *but* was presented with two frames—natural and Tone 2 lowered. The 11 trials were presented in random order. Listeners identified the target on each trial by marking a sheet on which the words *bit bet bat but* appeared. They were instructed to guess if they had no clear phonetic impression of the target.

The third segment of the test session was another test using sinewave signals and is not reported here.

The last segment of the session consisted of a 40-trial identification test in which the target syllables were presented 10 times each in random order for labeling in a four-alternative forced-choice paradigm. On each trial, subjects reported their impressions on a response form containing the words *bit bet bat but.* This test measured the distinctiveness of the target syllables independent of the influence of the framing sentences and served as a method for determining which subjects were unable to identify the sinusoidal targets consistently even under highly favorable conditions.

Subjects

Ninety-five undergraduate volunteers took the listening tests. None reported a history of speech or hearing disorder. Most were paid; others received course credit in Introductory Psychology for participating. All of the subjects were naive with respect to sinusoidal replicas of speech signals. They were tested in groups of two to six at a time, in visually isolated listening stations.

Results

Identification Test

In this test, the targets were presented in isolation for labeling, thereby determining the stability of the vowel categories given

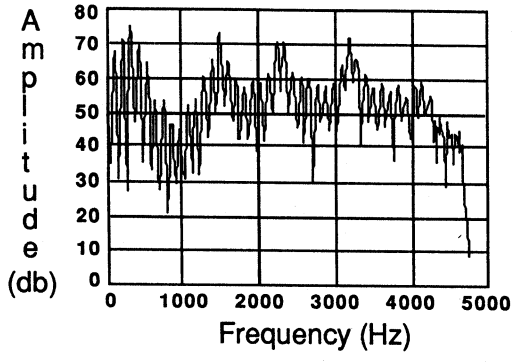

Figure 2. Spectral section of natural sentence.

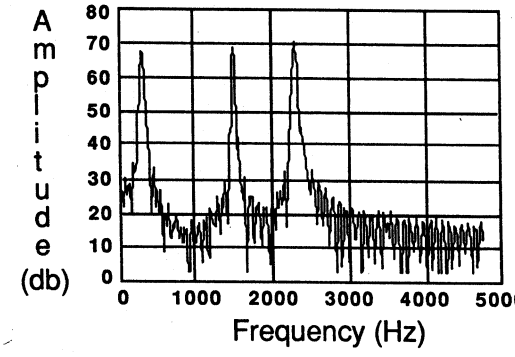

Figure 4. Spectral section of sinewave signal.

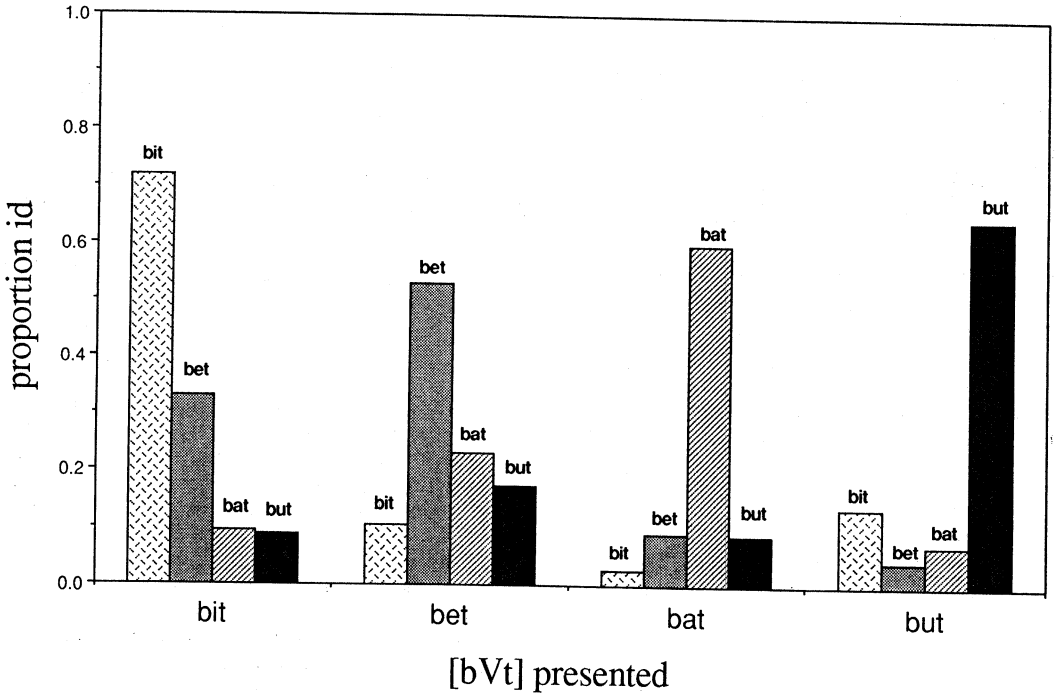

Figure 5. Group identification performance, [bVt] test, for all 95 subjects. (Proportion id = proportion of identification judgments.)

the sinusoidal presentation while also providing an index of the ability of each subject to treat sinusoidal signals phonetically. To expose both aspects of the data, the results of this test are presented in Figures 5, 6, and 7.

The first panel shows the identification performance for the entire group of 95 subjects. For each of the four targets, the distribution of labeling judgments is shown, the height of each bar within the foursomes representing the proportion of the responses assigned by the listeners to that category. Each target was presented 10 times; hence, each group of four bars represents 950 trials. The majority of responses agreed with the intended identity of the target, showing that brief [bVt] syllables are identifiable despite the sinusoidal realization of the resonance pattern. However, there is also evidence of inconsistent identification observable in the responses distributed across the unintended and less preferred alternatives. To resolve the outcome of this test more clearly, we have separately plotted the results contributed by those subjects whose performance is less differentiated overall from the results of those who appear more able to perform the identification task. Figures 6 and 7 portray the identification results for the listeners grouped in two sets: in Figure 6 those subjects who appeared relatively less able to identify sinewave patterns, and in Figure 7 the others who exhibited greater consistency.

The measure that we used to divide the subjects into two groups—those who could attribute phonetic properties to sinusoidal patterns reliably and those who could not—was straightforward. Because the preferred response alternatives of the entire group of 95 subjects matched the intended identity of the sinusoidal syllables, we determined for each subject the percentage of responses (out of the 40 trials) that fit this pattern. But we adopted a rather lenient cutoff, 50% of the available responses, as the threshold for identifying those subjects who simply failed to perform, thereby including with the majority those subjects whose performance may have been phonetically stable only for some of the syllables. In each instance, if fewer than half of a subject's responses were properly assigned to the intended categories across all four target syllables, then we designated that subject as unable to perform the task. Figure 6 shows the response distribution for the group of 33 subjects who selected the less preferred and unintended alternatives more than 50% of the time. Our examination of the individual subject data revealed rather little about this group, which evidently used a mixture of gambits including random labeling and arbitrary response preferences.

Figure 7 presents the identification data obtained from the remaining 62 subjects who assigned their responses to the majority category on more than 50% of the target presentations.

Table 1
Frequency Values of the Vowel Nuclei of the Three-Tone Target Syllables

Syllable	Tone 1	Tone 2	Tone 3	Duration
bit	393	1801	2572	110
bet	624	1688	2486	130
bat	805	1489	2376	160
but	624	1408	2477	150

Note. Frequency values in Hertz; duration in milliseconds.

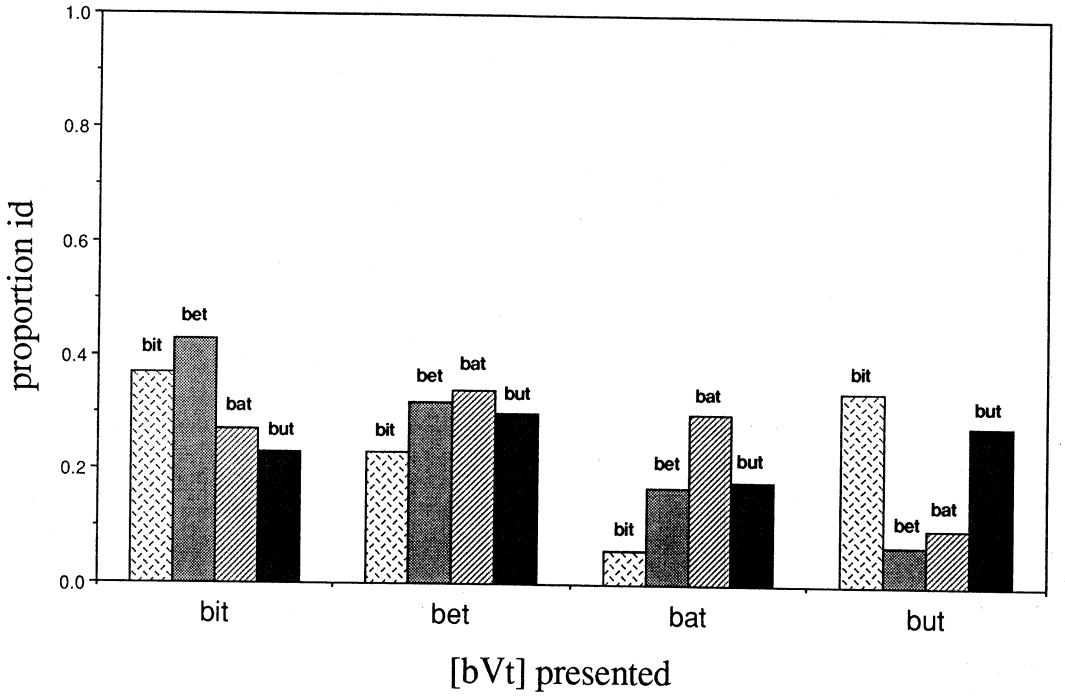

Figure 6. Group identification performance, [bVt] test, for 33 subjects whose performance was inconsistent. (Proportion id = proportion of identification judgments.)

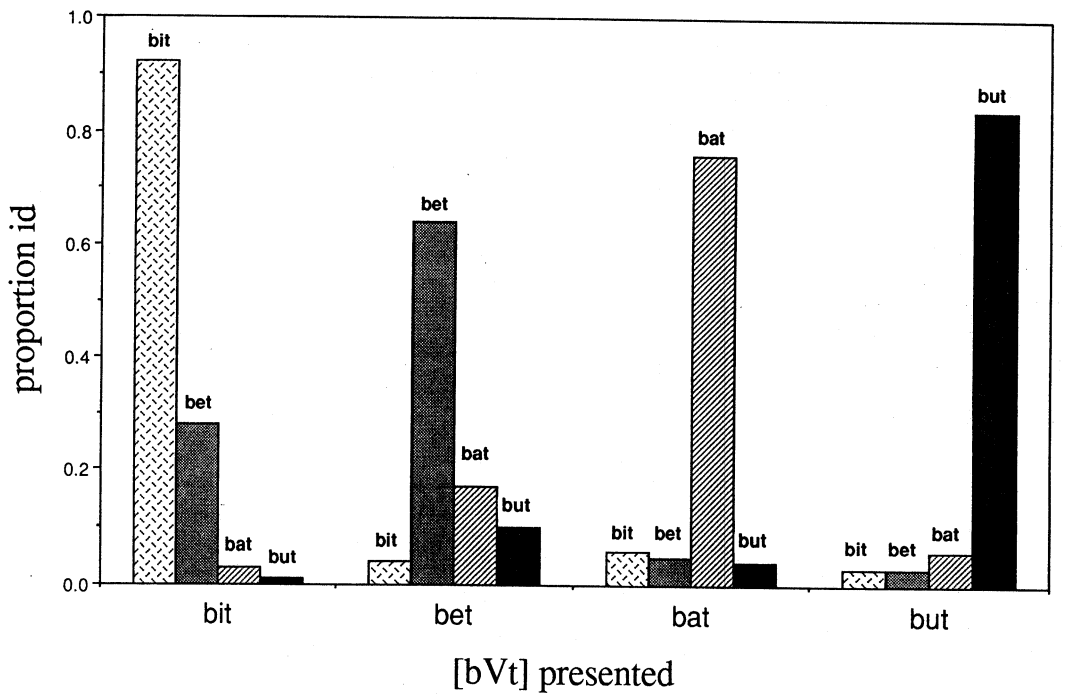

Figure 7. Group identification performance, [bVt] test, for 62 subjects whose performance was consistent. (Proportion id = proportion of identification judgments.)

This group is clearly able to attach vowel labels to sinusoidal patterns.

Overall, the data reveal that sinusoidal versions of [bVt] syllables contain sufficient variation to permit listeners to categorize the vowels. In addition, as many as one third of the subjects were found to be incapable of performing the labeling task, even under these conditions of low uncertainty, echoing the findings of Bailey et al. (1977), Best et al. (1981) Cutting (1974), and Williams et al. (1983).

Normalization Test

Because the identification test served as an independent index of each subject's susceptibility to phonetic perception with brief sinusoidal signals, we consider the normalization test results separately for the two groups of listeners whose performance is differentiated by this test. The null hypothesis for each of the four target conditions is the same: If normalization did not occur, then identification of the target will be indifferent to the properties of the precursor sentences, and the distribution of the identification judgments will not be altered by different precursor sentences. In fact, to summarize the results, the test showed the influence of the leading sentence in three of the four target conditions with phonetic listeners, and in one of the four conditions with nonphonetic listeners.

Phonetic listeners. The results for this group of 62 listeners are shown in Figures 8–11. When the target syllable was *bit,* shown in Figure 8, no difference was observed in labeling for the two sentence frames: the first natural (the sentence with unmodified frequency values), the second with Tone 1 lowered, $\chi^2(3, N = 62) = 3.62, p > .25$. This differs from the finding reported by Ladefoged and Broadbent, in whose study this modification of the framing sentence caused the target syllable *bit* to appear to be *bet.*

As shown in Figure 9, the target syllable *bet* was presented with four different sentence precursors. Ladefoged and Broadbent had found that one of these, in which the first formant was raised, created the impression that the target was *bit.* Similarly, we found that the analogous transposition of the sinusoidal precursor sentence increased the proportion of *bit* responses to the *bet* target. This effect was significant, $\chi^2(9, N = 62) = 20.72, p < .025$.

Figure 10 shows the outcomes for the three precursor sentences that were used with the target *bat.* Ladefoged and Broadbent had found that raising the first formant in the precursor sentence alone altered the perception of the target, increasing the proportion of trials on which *bat* was identified as *bet.* We also observed the same effect, in which instance raising Tone 1 of the sentence increased the frequency of *bet* labels applied to the *bat* target. In addition, raising Tone 2 made identification difficult, with the single exception that subjects consistently rejected *but* as the identity of the target. These differences attributable to the precursor sentences are statistically significant, $\chi^2(6, N = 62) = 28.35, p < .001$.

In the last condition, portrayed in Figure 11, the target *but* was presented with two precursors, the natural sentence and a sentence with Tone 2 lowered. The difference in labeling attributable to the precursor is significant, $\chi^2(3, N = 62) = 13.42, p < .005$, though the pattern of outcomes is unlike the effects reported by Ladefoged and Broadbent. They observed a shift in responses from *but* to *bat,* while we saw a shift from *but* to *bet.* Nonetheless, in this condition we may again identify the influence of the leading sinusoidal sentence on the vowel of the lagging target syllable.

Nonphonetic listeners. The results of the four target syllable conditions for the 33 nonphonetic listeners designated by the identification test are shown in Figure 12. In three conditions, no statistical differences were observed in the response distributions for different precursor sentences. When *bit* was the target (Figure 12), $\chi^2(3, N = 33) = 2.35, p > .5$; when *bet* was the target (Figure 13), $\chi^2(9, N = 33) = 6.06, p > .5$; when *but* was the target (Figure 15), $\chi^2(3, N = 33) = 1.91, p > .5$.

When the target was *bat,* it was poorly identified with the natural precursor; it was identified as *bet* when the precursor contained a raised Tone 1; and it was identified as *bit* when the precursor contained a raised Tone 2, $\chi^2(6, N = 33) = 23.86, p < .001$. Note that this pattern differs greatly from what we observed for the phonetic listeners (compare Figure 10 and Figure 14) and is also unlike the outcome observed by Ladefoged and Broadbent in the analogous condition.

Discussion

Overall, these data reveal a pattern of vocal-tract scale effects in three-tone vowel identification. This implicit accommodation for variation in vocal-tract dimensions occurs despite the fact that the phonetic and vocal-tract information is carried by an anomalous sinusoidal signal. Although there are a few points of dissimilarity between the present data and those described by Ladefoged and Broadbent, the results are similar. They may therefore be taken here as evidence for talker normalization and, hence, for perceptual evaluation of sinusoidal replicas akin to ordinary perception of speech, and for the likely reliance of this process on the time-varying information in sinusoids ordinarily present in natural signals. Accordingly, postperceptual guesswork or strategic problem solving is probably not primary in the transcription of tonal patterns, though it cannot be eliminated as a secondary contribution.[1] To offer an account of the outcome, though, we must address three prominent issues raised by the results: (a) In detail, the results for phonetic listeners in this study differ from the oucomes reported with synthetic speech; (b) despite the apparent inability to identify the targets, the nonphonetic listeners were influenced by the precursor sentences in the instance of the *bat* target; and (c) most generally, our test seems to differentiate phonetic and nonphonetic listeners, though it is unclear what this means.

A Comparison With Ladefoged and Broadbent

The effect of normalization can be likened to recalibration of the formant frequency values of the perceptual standards for vowel categories. Ladefoged and Broadbent took this approach. Presumably, the changes in target identity that we observed in the sinusoidal cases likewise were caused by resetting the standards for vowel perception, all other things being equal (see

(*Text continues on p. 51*)

[1] Postperceptual guesswork and strategic problem solving also play a secondary role in ordinary perception of speech.

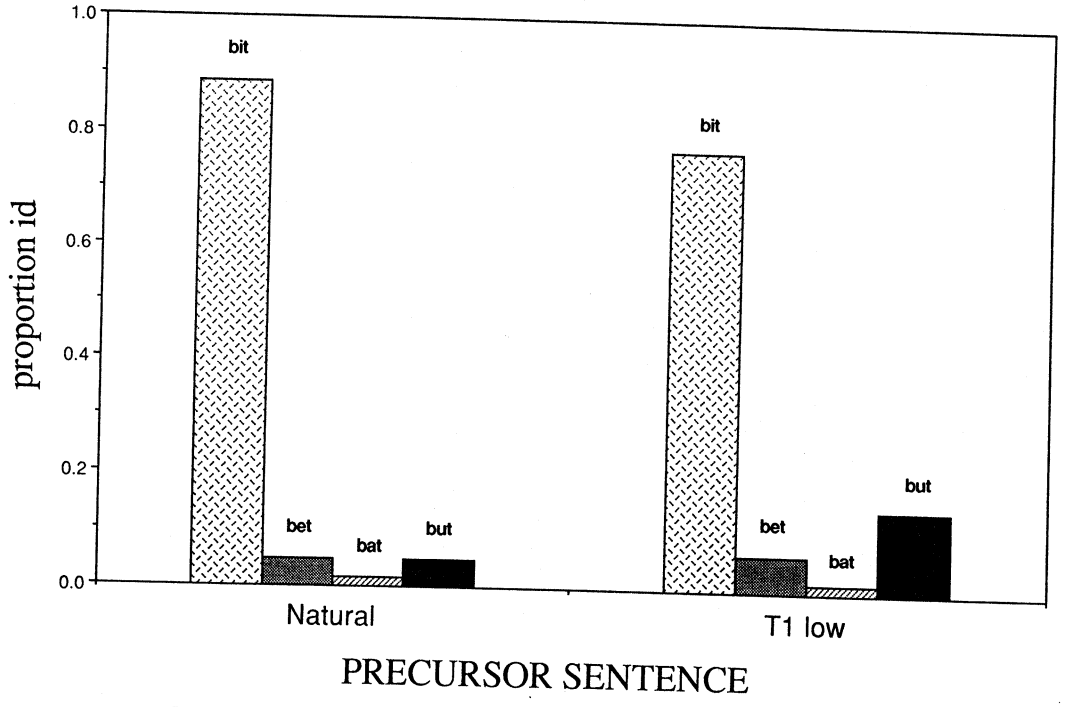

Figure 8. Normalization test performance, phonetic listeners: conditions with *bit* target. (Proportion id = proportion of identification judgments.)

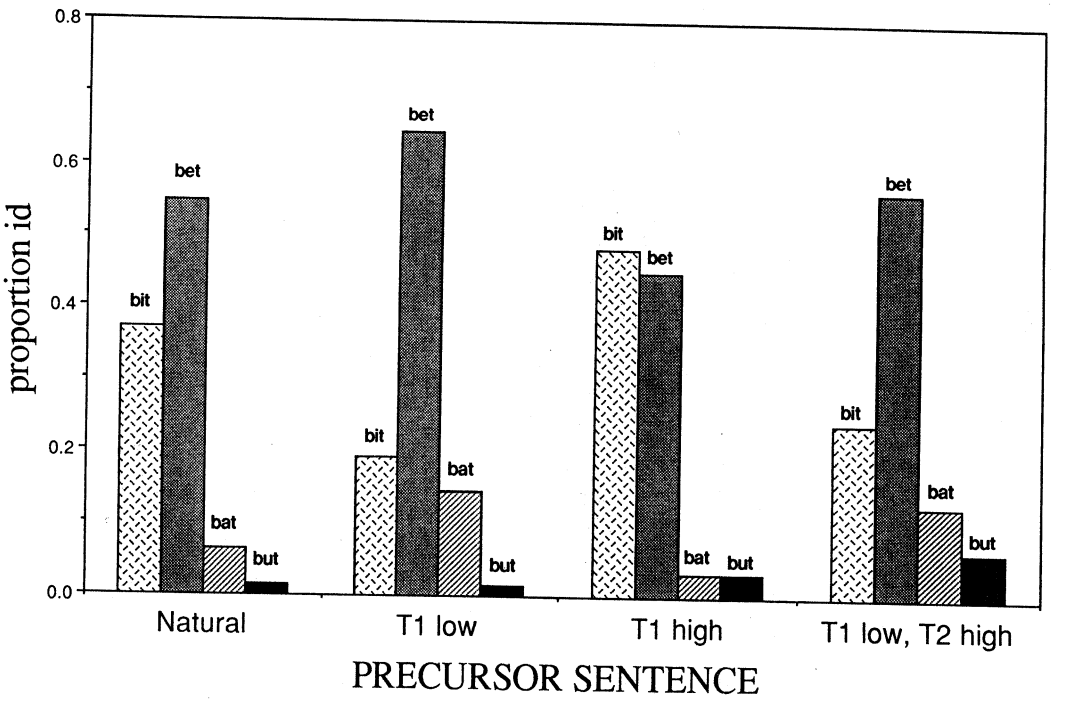

Figure 9. Normalization test performance, phonetic listeners: conditions with *bet* target. (Proportion id = proportion of identification judgments.)

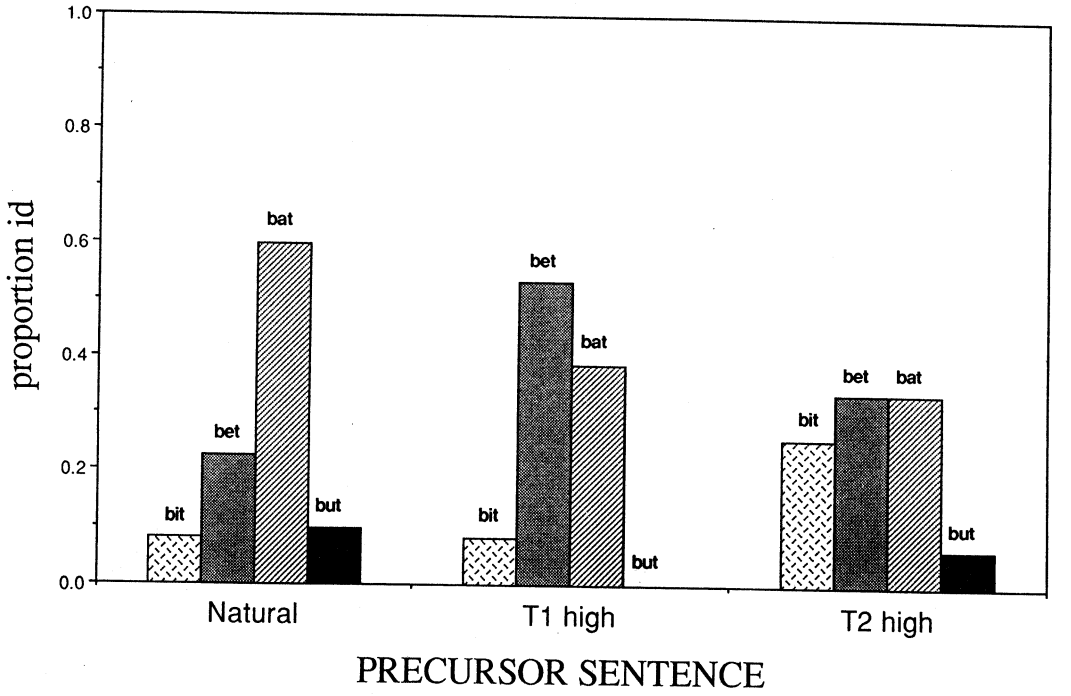

Figure 10. Normalization test performance, phonetic listeners: conditions with *bat* target. (Proportion id = proportion of identification judgments.)

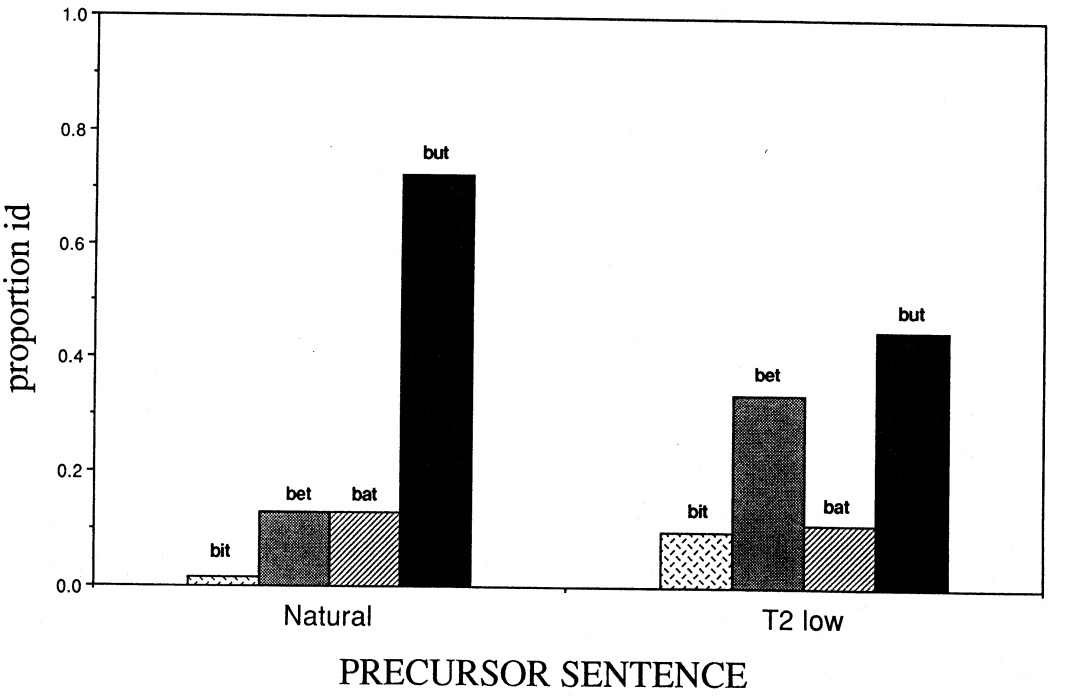

Figure 11. Normalization test performance, phonetic listeners: conditions with *but* target. (Proportion id = proportion of identification judgments.)

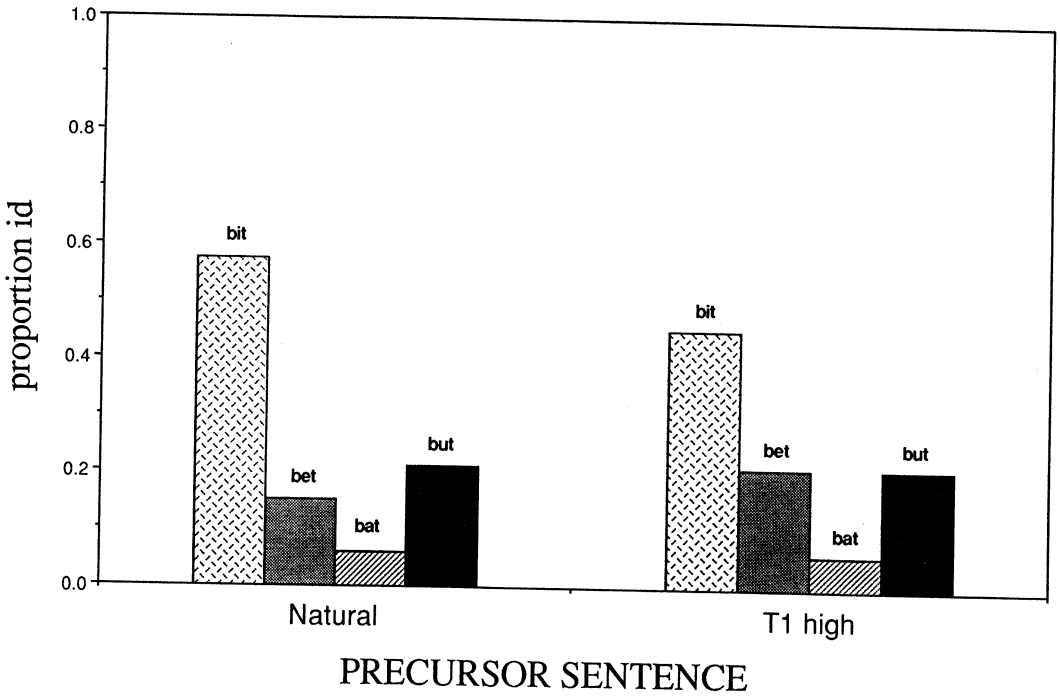

Figure 12. Normalization test performance, nonphonetic listeners: conditions with *bit* target. (Proportion id = proportion of identification judgments.)

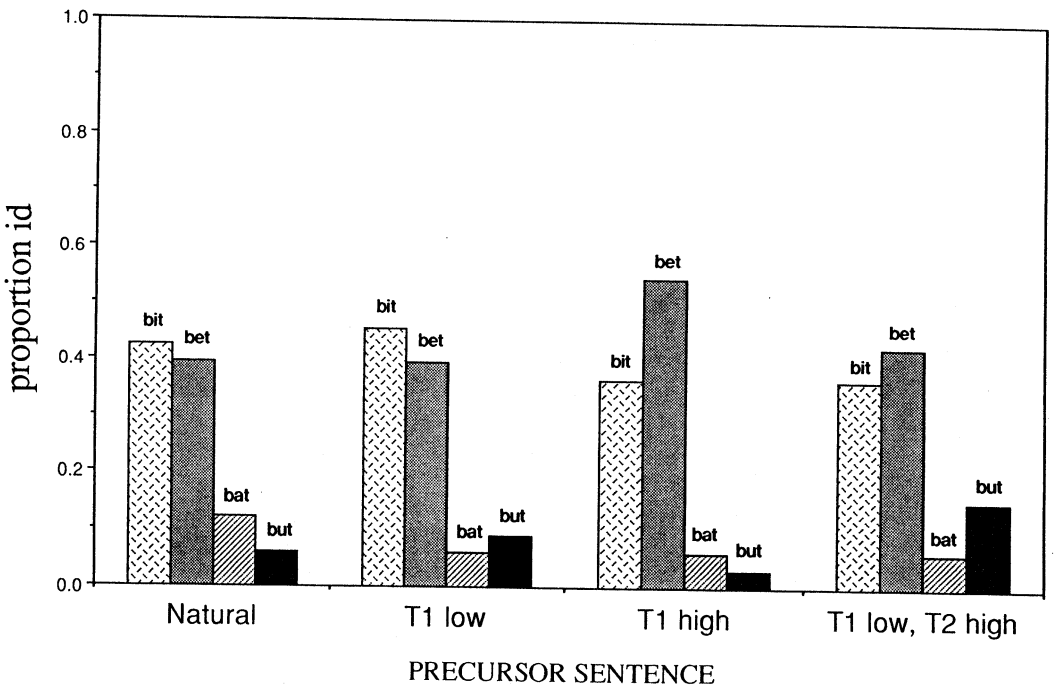

Figure 13. Normalization test performance, nonphonetic listeners: conditions with *bet* target. (Proportion id = proportion of identification judgments.)

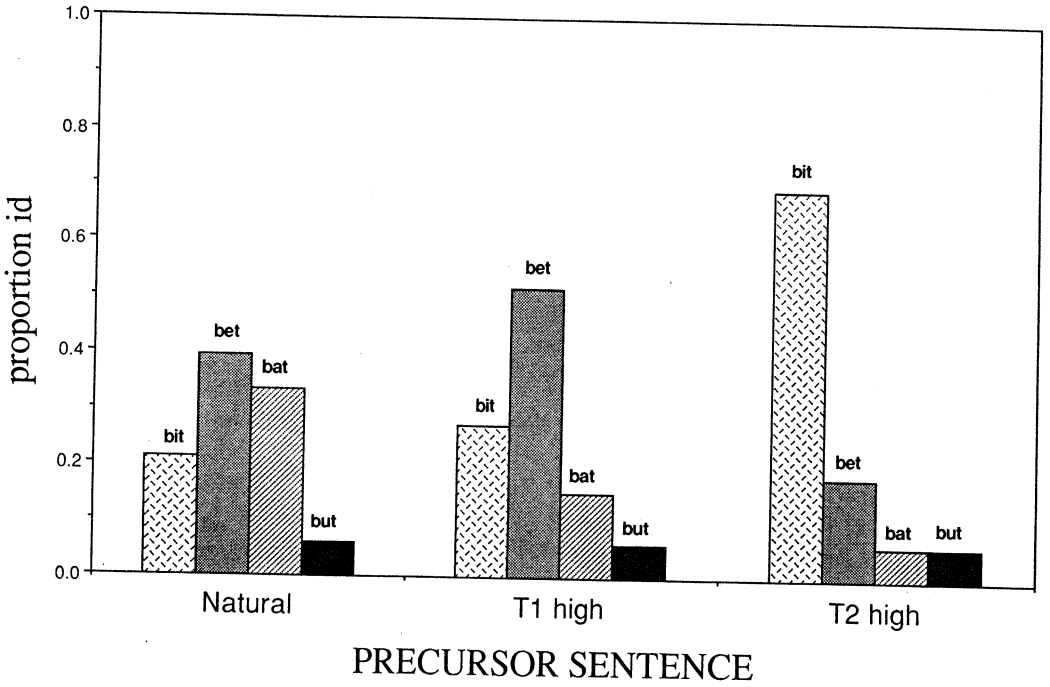

Figure 14. Normalization test performance, nonphonetic listeners: conditions with *bat* target. (Proportion id = proportion of identification judgments.)

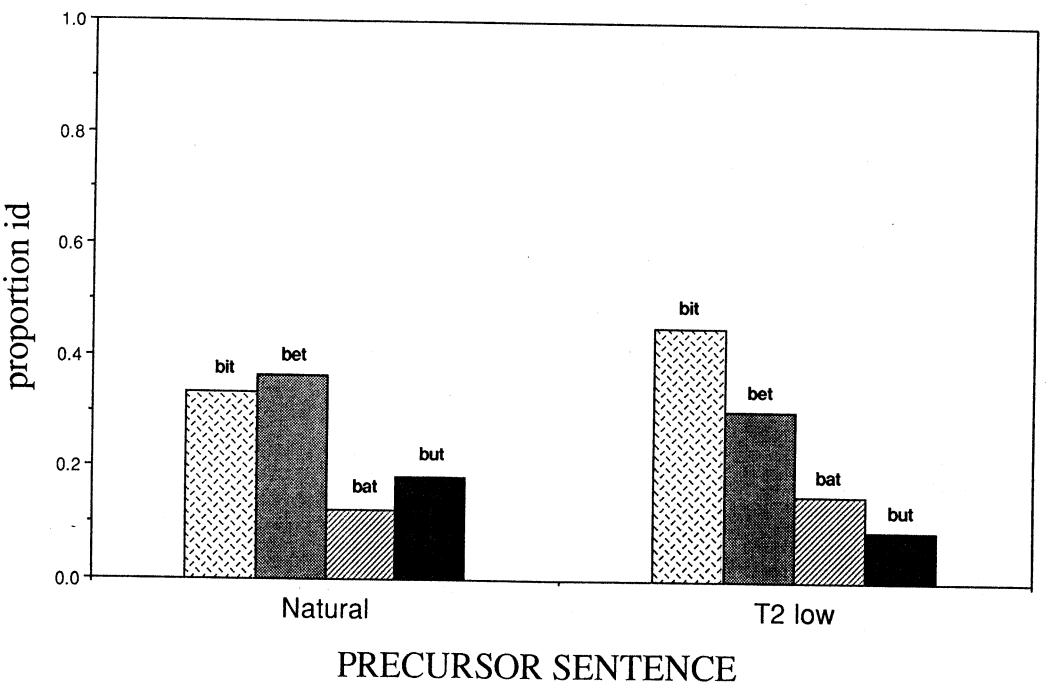

Figure 15. Normalization test performance, nonphonetic listeners: conditions with *but* target. (Proportion id = proportion of identification judgments.)

Ainsworth, 1974, for a discussion of a durational component in normalization). To draw a parallel to Ladefoged and Broadbent's precedent, we adopt their heuristic, defining the perceptual standard for a vowel as a point in two dimensions, frequency of Tone 1 (the analog of the first formant) by frequency of Tone 2 (the analog of the second formant).

Operationally, we assume that the first and second formant frequencies of the syllable nuclei given in Table 1 approximate the vowel standards for the vocal tract that issued the natural sentence and the four target syllables. The effects of normalization can be estimated by transposing the vowel standards in accordance with the various precursor sentences that we used. Then, projecting the untransposed tone frequencies of the targets against recalibrated perceptual standards permits the evaluation of the target within a normalized vowel space.

Four examples of this approach are portrayed in Figures 16–20. We found that when Tone 1 was lowered in the precursor sentence, neither the categorization of *bit* nor *bet* changed. (Following the procedure of Ladefoged & Broadbent, the two other sinusoidal targets were not tested with this precursor.) In Figure 17, we represent the effect of recalibrating the vowel standards by lowering the Tone 1 value by 25%, which places the *bit* and *bet* targets in positions between the putative category centers rather than within new categories. This simplification of normalization agrees with the observations in the sinewave cases. In contrast, Ladefoged and Broadbent had observed *bit* identified as *bet* and an absence of a normalization effect in the case of the synthetic *bet* target. Our sinewave results appear consistent with the rationale, if not the specific finding, of the earlier investigation. There is a probable reason why this occurred. The sinewave targets and sentences were linguistically similar to the synthetic speech of the prior study but were not acoustically identical in the values of spectral peaks, either in the sentence set or in the targets. Our test materials also differed from the original in the dialect of English that was used. In the present test, sinewave synthesis values were derived from new samples of natural speech and therefore replicated natural productions with the approximate linguistic and acoustic attributes, rather than replicating in detail the synthetic acoustic materials of the original test. In consideration of this, a departure from the fine grain, though perhaps not the general finding, of the earlier research is to be expected. The interpretation ultimately hinges on the internal consistency of the present findings and the general correspondence of the sinewave results to the synthetic speech precedent. In this respect, the results of lowering Tone 1 in the sentence frame are best considered along with the outcomes of other conditions.

Raising Tone 1 in the precursor sentence rendered the sinewave *bet* ambiguous, with responses mostly divided between *bet* and *bit,* and made *bat* seem like *bet.* As Figure 18 reveals, the normalized *bet* target in this instance is intermediate between the IH and EH vowel categories, while the *bat* target falls between EH and UH. Because our listeners judged *bat* to be like *bet* on 20% of the natural context trials, perhaps the intermediacy of the *bat* target between EH and UH should be interpreted as an advantage for EH. A similar argument may apply to the *bet* target in this condition with respect to IH and EH. In both cases, the figure minimally suggests that raising Tone 1 should make the *bet* and *bat* target vowels seem higher along the vowel dimension high–low, and this is a reasonable description of the performance of our listeners. Ladefoged and Broadbent, in conditions that ours paralleled, observed that raising the frequency of the first formant in the precursor sentence caused *bet* to be identified as *bit* and an ambiguous *bat* to be identified as *bet,* both changes in vowel height. In other words, the effects we observed have the same pattern as the synthetic precedent, though they fall short of complete parity because of acoustic differences between the sinewave and synthetic materials, because the variability in the sinewave response data partly obscures the evidence of normalization, or because the transpositions that we imposed on the precursor sentence were too moderate.

Lowering the frequency of the second tone of the vowel standards brings the *but* target into the EH category, depicted in Figure 19. This approximately matches what our listeners did. In the original synthetic version, though, Ladefoged and Broadbent's listeners had judged *but* as ambiguous with a similar frequency transposition of the sentence frame, half of the judgments for *but* and half for *bat.* From Figure 19, though, it is clear that this sinusoidal condition should project the *but* target into the EH category in the present case.

The precursor with Tone 2 raised was presented with the *bat* target only, in which case we found that listeners apparently could not identify the target, with the exception that they consistently rejected *but* as the target identity. This sentence, then, made the subjects perceive the *bat* target no more precisely than as "not *but.*" Figure 20 reveals the normalized standards to be quite remote from the values of the *bat* target, though the exclusion of UH from the possibilities cannot be derived from this representation. Ladefoged and Broadbent found that this transposition of the precursor sentence had no effect on the identification of the *bat* target, though even in their natural context this target was less consistently identified than the other three, and, incidentally, less consistently identified than the sinewave *bat* in our study.

To summarize, the results of the sinewave test appear similar to the original synthetic precedent, and the discrepancies may be accounted for by much the same argument that Ladefoged and Broadbent employed to explain their observations. Essentially, the results obtained with synthetic speech and sinewave replicas agree. In consequence of this, the motivation of the explanation for the synthetic results is no less applicable to the sinusoidal results—that is, for the listeners who perceive sinewaves phonetically. Because this is evidence of normalization specific to the acoustic correlates of vocal dimensions, it is fair to say that sinewave replicas of speech signals and ordinary speech enjoy a common perceptual treatment.

Nonphonetic Influence on the Bat Target

The identification test had designated a group of listeners who exhibited nonphonetic perception of sinusoidal replicas of speech, and these subjects performed the normalization test in a manner largely indifferent to the range of the tone variation of the precursor sentences. However, they did exhibit significantly different distributions of identification judgments across the three precursors used with the *bat* target. The performance on this condition by nonphonetic listeners differed dramatically from that of the phonetic listeners, as Figures 10 and 14 show,

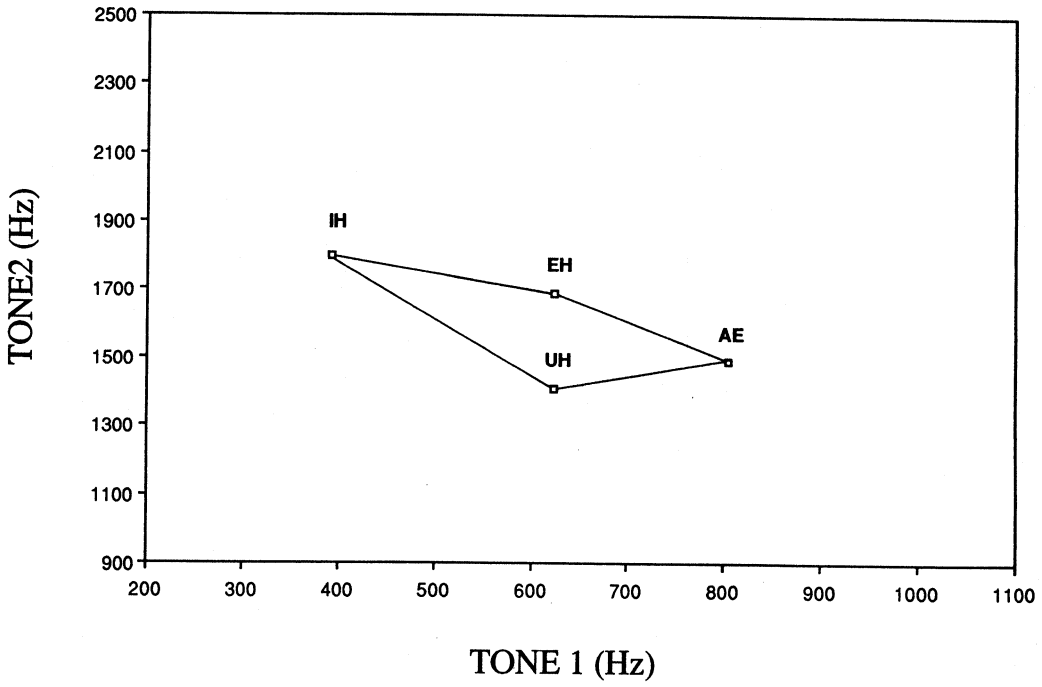

Figure 16. Sinewave nuclei: normalization portrayed as recalibration of perceptual vowel standards. (Hypothetical values for normalized vowel standards are estimated by transposing the first or second formant frequencies observed at the syllable nuclei of the four target vowels, and are labeled IH, EH, AE, and UH. Within each condition, the appropriate target syllable is represented as a point in this space, and is labeled in the format bVt. This figure shows the relative positions of untransposed targets.)

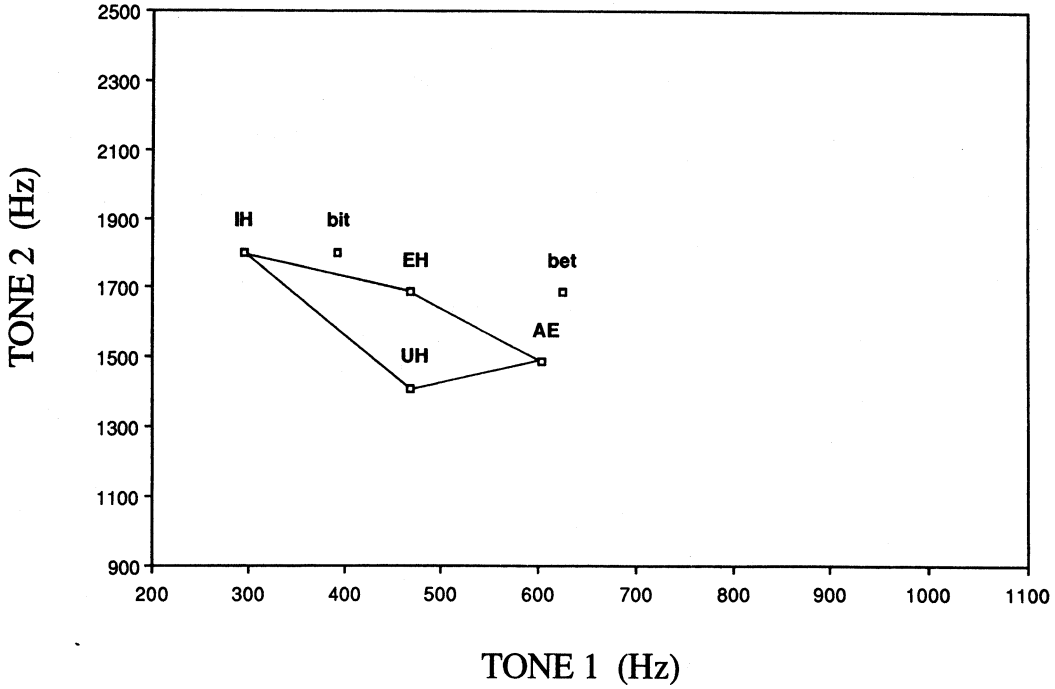

Figure 17. Sinewave nuclei: normalization portrayed as recalibration of perceptual vowel standards—effect of lowering first formant tone by 25%. (Normalized standards = IH EH AE UH targets = bVt.)

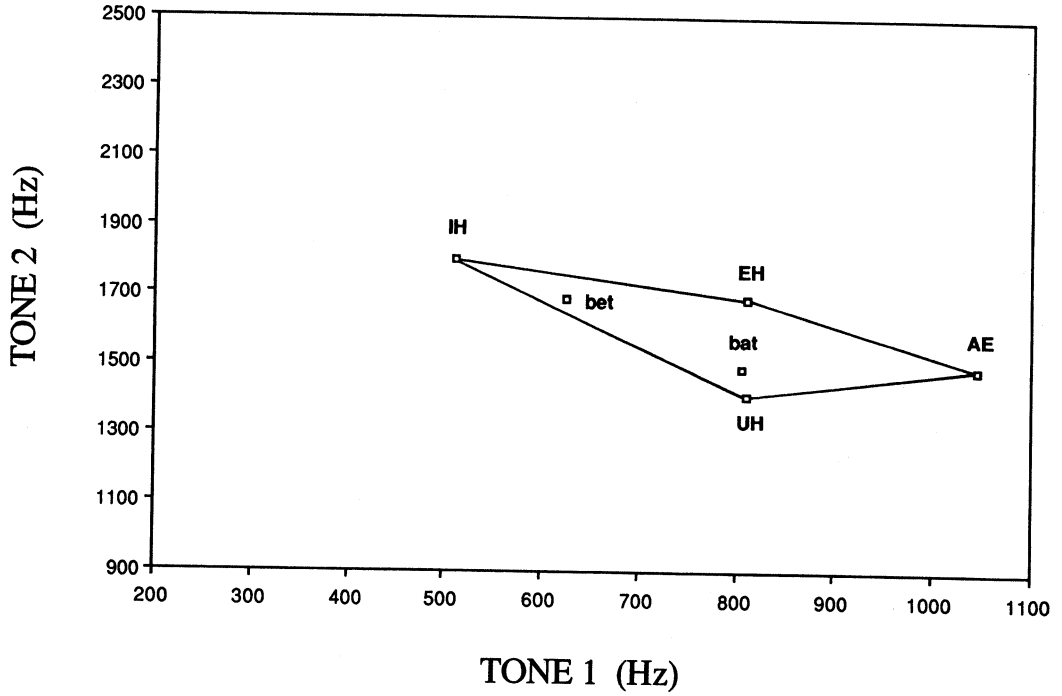

Figure 18. Sinewave nuclei: normalization portrayed as recalibration of perceptual vowel standards—effect of raising first formant tone by 30%. (Normalized standards = IH EH AE UH targets = bVt.)

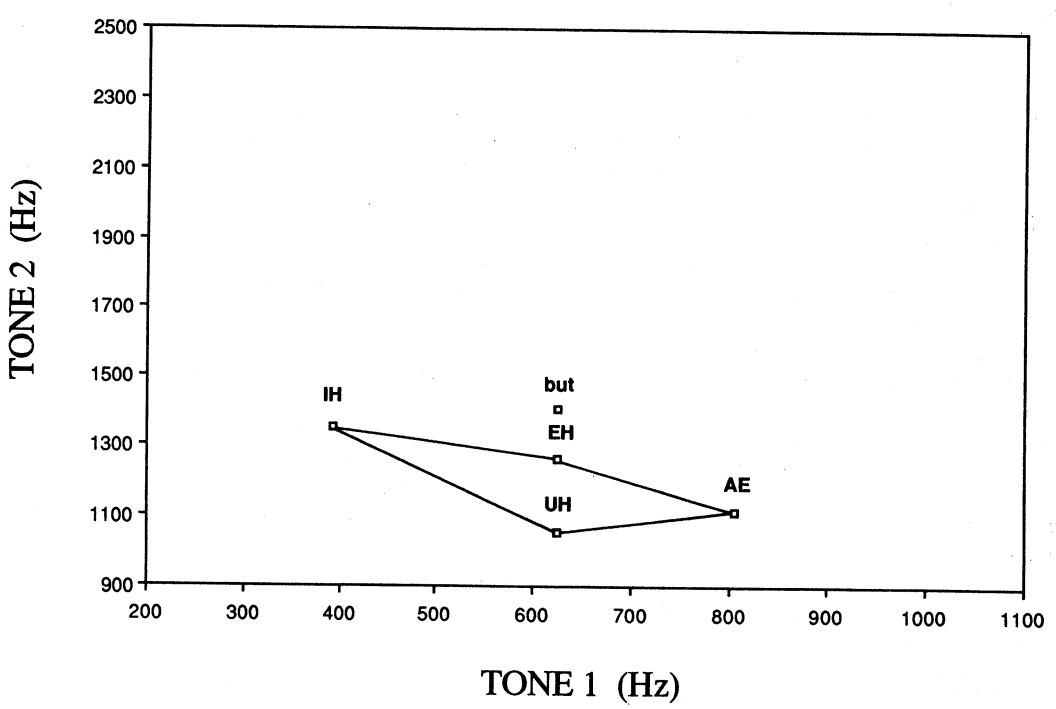

Figure 19. Sinewave nuclei: normalization portrayed as recalibration of perceptual vowel standards—effect of lowering second formant tone by 25%. (Normalized standards = IH EH AE UH targets = bVt.)

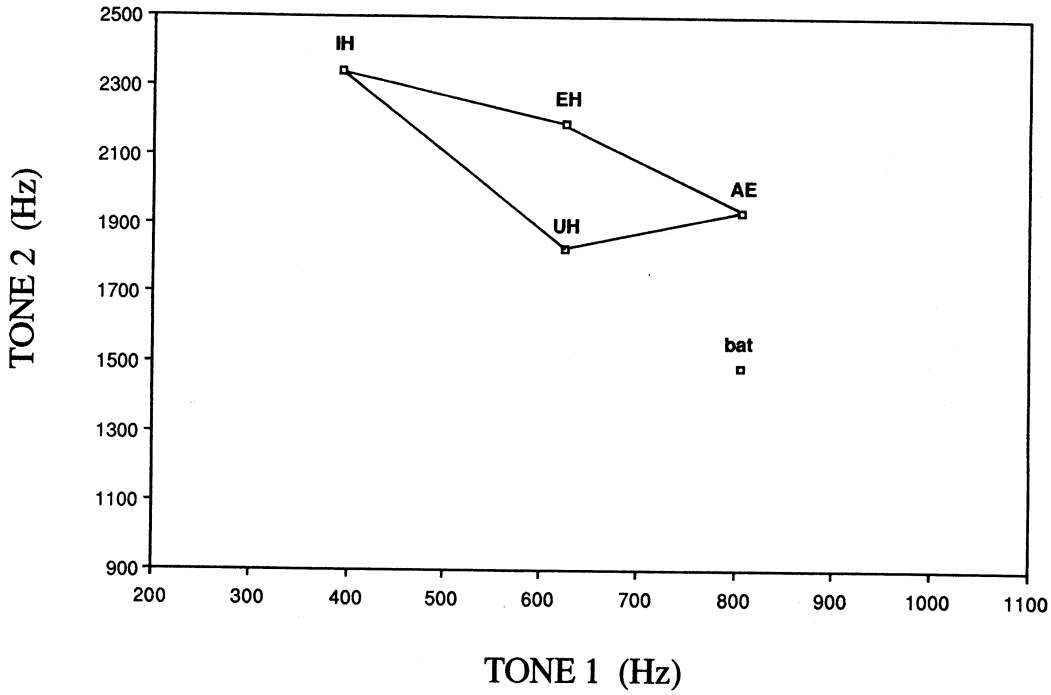

Figure 20. Sinewave nuclei: normalization portrayed as recalibration of perceptual vowel standards—effect of raising second formant tone by 30%. (Normalized standards = IH EH AE UH targets = bVt. The effect of the precursor sentence transposed by simultaneous lowering of Tone 1 and raising of Tone 2 is not shown in this set of figures. This condition was effective perceptually neither in the present study nor in the study by Ladefoged & Broadbent, 1957.)

which offers a hint at the cause of this outcome. The significant nonphonetic effect may be traced to one detail: The effect of raising Tone 2 in the precursor elicited consistent identification of the target as *bit*. This appears not to be an intrinsically phonetic effect and is completely at odds with the rationale that Ladefoged and Broadbent developed and that we have applied here. Raising Tone 2 places *bat* farthest from the IH standard of the four possible choices, as Figure 20 shows. It is the least likely identification, according to the phonetic rationale. Although the performance of this group attained statistical significance in this target condition, it is unlikely to have arisen due to selective susceptibility to normalization or to phonetic properties of sinusoidal signals when the vowel in the target is AE. This speculation is far from definitive, though, and much remains to be clarified about the performance of subjects who are unable to attain phonetic organization of sinusoidal replicas of speech.

Phonetic and Nonphonetic Listeners

Our prior investigations of the perceptual attributes of sinusoidal signals led us to expect that no less than a third of the subjects we encountered would be incapable of characterizing the signals phonetically. Alternative modes of perceptual organization appear to underlie this circumstance, one a phonetic organization that takes multiple tone variation to be a single fused pattern, and another that takes the tones each to be an independent stream. An account of this difference and the definition of the effective trigger for coherent organization will require additional study. In practice, though, this finding exemplifies an ongoing difficulty with tests of perceptual sensitivity to time-varying information. Because receptivity to sinusoidal messages is not easily predicted, we used the test of identification here to index each subject's capability. This use of a converging identification test separated two groups whose performance on the normalization test differed considerably and therefore appears to be an objective way of responding to the methodological challenge posed by the sinusoidal replication technique.

Other researchers who have conceptualized their sinewave results similarly, as the product of two (or more) perceptually distinct populations, have occasionally—and to good effect—used less formal post hoc practices to identify a subject's membership in one group or the other (Bailey et al., 1977; Best et al., 1981); our a priori approach here is surely not less suited to the problem. A theoretically richer account of individual differences in speech perception, including the issue of perceptual organization that concerns us here, may eventually be pursued with the sinusoidal replication technique, though it is beyond the scope of this study.

Conclusion

Because the outcome of our test using sinusoidal replicas of speech resembles the outcome of the test employing conventional synthetic speech, then one possible interpretation is that

the two instances share a common perceptual treatment. We favor this conclusion and the implication that sinusoidal signals preserve information provided by coherent spectral changes in ordinary speech signals. However, an alternative to this perspective would be to view this experiment as a falsification of the classic study after which it was modeled. It could be argued that the replication of rescaling with nonspeech sounds invalidates the account given originally by Ladefoged and Broadbent, which alleged that the effect was based on the personal information available in vocal sound production. In fact, there have been challenges to the account given by Ladefoged and Broadbent pertaining to the role, or the sufficiency, of resonant spectral information in perceptual scaling and vowel perception (e.g., Ainsworth, 1974; Dechovitz, 1977b; Ladefoged, cited in Papçun, 1980; Syrdal & Gopal, 1985; Thompson & Hollien, 1970; Verbrugge et al., 1976). Although these reports proposed technical improvements, variations, and extensions of the initial paradigm, warranting the expansion of the proposed mechanism for normalization, the rationale of normalization has proven to be quite durable (see Nearey, 1978).[2] Although the precise perceptual mechanism involved may be obscure, our finding looks like range-appropriate perceptual rescaling, and we therefore appeal to the conventional explanation—vocal-tract normalization.

Looking across the results of Experiment 1, we find that the sinewave outcomes approximated the findings of the synthetic speech version serving as our theoretical and procedural model. The identification of a target word was evidently affected by the variation of tones within the preceding sentence. This was formulated by Ladefoged and Broadbent as a contingency of phonetic perception on personal information about the talker, and we propose to extend this explanation to the present finding. To state the conclusion, the listener perceives sinewave replicas of speech by implicitly recognizing that the tonal complexes originate from a vocal source. Accordingly, sinewave signals appear to be handled by the ordinary means of speech perception, probably because such abstracted spectral patterns nonetheless preserve time-varying phonetic information present in natural signals.

Experiment 2

Control Condition About Vowel Perception

The first experiment of this report tested the assumption that sinusoidal signals are perceived phonetically, using a measure that reflected the contingency of vowel identification on perceptual organization. Evidently, sinusoidal patterns are identified to be vocal in origin, despite the absence of the acoustic products of vocalization in the signal. In consequence, perceptual normalization of vocal-tract dimensions influenced the identification of the vowels in the target syllables, as happens in instances of synthetic and natural speech.

A prominent component of accounts of vowel perception more generally is, simply, that the frequencies of the first and second formants together provide much of the information for vowels in English (reviewed by Ladefoged, 1967; Nearey, 1978). With few exceptions (e.g., Rubin, 1971; Shepard, 1972), these two dimensions of acoustic variation, first and second formant frequencies, have offered a compelling characterization of the perceptually significant acoustic correlates of vowels, especially so because they are held to correspond to articulatory dimensions of sound production. Though we must take the anatomical designations loosely, it is as if the first formant were associated with the height of the tongue, high or low, and the second formant with the advancement of the tongue, front or back.

The sinewave syllables used as targets in Experiment 1 present spectral patterns derived from frequency variation of these acoustic sources of vowel information for the perceiver. Nevertheless, the possibility exists that listeners identified the sinewave vowels in a manner unlike ordinary vowel perception, due to the unspeechlike spectrum that such signals present to the ear. This is encouraged by reports that listeners can attribute speechlike qualities to simple acoustic signals. Of course, it is common to assign a phonetic label to a nonspeech impression in instances of onomatopoeia, though this is presumably distinct from ordinary speech perception, at least in the respect that nonspeech impressions of formant variation do not typically accompany phonetic perception. In the case of sinewave replicas, though, the nonspeech quality of the signal persists even when transcription of linguistic properties occurs (Remez et al., 1981). In order to conclude that the influence of precursor sentences on target identification, observed in Experiment 1, was evidence of vowel normalization, it will be necessary to establish more firmly that the perception of the vowels in the targets was an instance of vowel perception, differing from the attribution of similarity between the nonspeech auditory qualities the listener hears and the vowel features the listener knows.

Vocality

Psychoacoustic investigations of an earlier generation examined the basic sensory attributes of auditory experience, among which are found the familiar pitch, loudness, and timbre. Additionally, simple acoustic presentations were held to cause the experience of *vocality,* a kind of speechlike quality of sound that was irreducible to simpler experience (Boring, 1942; Gatewood, 1920; Köhler, 1910; Modell & Rich, 1915). Operationally, vocality was observed in studies in which the ordinary acoustic dimension of frequency was varied, but the subject responded with a vocal imitation or a phonetic segment name instead of reporting a pitch experience. For example, this sequence of sensations of vocality was held to occur with ascending frequency from 65 Hz to 33.6 kHz:[3] "vvv-mmm-U-O-A-E-I-sss-fff-ch" (Boring, 1942; page 374).

Since these pioneering efforts, the acoustic correlates of pho-

[2] Suomi (1984) proposes a model of vowel identification that essentially filters out vocal-tract scale variation but that also ignores contingent effects of the kind that we reviewed here.

[3] The value given for the upper frequency bound of vocality impressions reflects a compound error (Boring, 1942). Vocality was said to vary across the dimension of frequency, but contemporary estimates of frequency were erroneous. Because of mistaken calibration of the Galton whistles that were used as acoustic sources, the upper frequency bound of audibility had been overestimated at 50 kHz. In consequence, the frequency limits for impressions of vocality were set well within the range of audibility, as it was understood at the time.

netic perception have been more realistically defined as complex spectra rather than as an accompaniment of pitch sensation. But for interpreting the present findings, the precedent of vocality studies is uniquely relevant. Specifically, if subjects are able to label nonspeech simple tones with vowel names (Fant, 1959; Modell & Rich, 1915), then the performance of subjects in Experiment 1 may have exploited this capacity for designating likeness rather than for perceiving vowels. Our prior research has shown that Tone 1 is especially prominent perceptually (Remez & Rubin, 1984), which suggests that subjects may have analyzed this tone as an independent component and may have based the vowel responses on the phonetic likeness of this tone. If our subjects differentiated the four vowel targets on the basis of the distant phonetic likeness of simple nonspeech spectra, rather than on information given in the multiple spectral properties to which most accounts of vowel perception refer, then a strong claim about speech perception from coherent spectral variation is inappropriate. An experimental investigation of this possibility is clearly warranted.

Experiment 2, then, is a test to distinguish the assignment of vowel names to perceptually prominent tones from the perception of sinusoidal vowels. The objective is to determine whether the identification of the vowel of the three-tone target syllables may be accounted for by vowel labels applied to individual component tones. To the extent that listeners report the same vowel from three-tone and single-tone patterns, we may doubt that vowel perception of sinusoidal syllables involves the ordinary mechanism of speech perception. To the extent that identification of single tone and multiple tone patterns differ, we may conclude that vocality judgments are a different sort of perceptual phenomenon than speech perception from sinusoidal replicas of natural utterances.

Several conditions were used in this attempt to distinguish sinusoidal vowel perception from the attribution of phonetic qualities to nonspeech signals. In the first, we simply replicated the identification test of Experiment 1, in which the four three-tone targets were presented to phonetically disposed listeners. Then, an identification test was presented consisting solely of Tone 1 from each of the four targets. A third test determined the identification of Tone 2 from each target presented in isolation. Last, two conditions investigated the perceptual interaction of spectral peak frequency and gross spectral shape by presenting Tone 1 and Tone 2 synthesized as sawtooth waves rather than as sinewaves. As an ensemble of conditions, these tests define the acoustic correlates of vowel judgments in Experiment 1, revealing whether or not the phonetic likeness of nonspeech impressions plays a role in sinusoidal phenomena.

Method

Acoustic Test Materials

Our design here called for five sets of tonal patterns. The first consisted of the four three-tone [bVt] target syllables of Experiment 1. The second set was composed of the lowest component of each of the four syllables, the Tone 1 set. The third set contained the four isolated Tone 2 sinusoids. The fourth and fifth sets consisted of single tone patterns, realized as time-varying sawtooth waves, in one case following the frequency values of the four Tone 1 patterns, in the other, the patterns of the Tone 2 set. Sawtooth waves presented the same frequency peak in the acoustic spectrum but possessed a shallow roll-off of the high-frequency skirt of the spectrum envelope. Test orders were compiled on the VAX and then output to audiotape, as in the first experiment of this report, and were delivered to listeners binaurally over headsets in the manner of Experiment 1.

Procedure

Following a warm-up sequence of eight sinusoidally synthesized sentences were three brief tests. First, a 40-trial identification sequence was presented in which each of the three-tone syllables occurred 10 times each in a random order. On each trial, subjects marked a response form, circling the word *bit, bet, bat,* or *but* in a four-alternative forced-choice paradigm.

Next, subjects were presented with an 80-trial test presenting the eight single-tone patterns of the Tone 1 and Tone 2 conditions. They occurred intermixed, 10 times each, in random order. Again, a subject chose one of the four alternative responses on each trial by circling the appropriate word in the test booklet.

Finally, a second 80-trial identification test, composed of the eight single sawtooth patterns, was presented. The eight patterns occurred 10 times each in random order. On each trial, a subject identified the word containing the vowel closest to the sound of the tone, marking the response form accordingly. The entire session, including the warm-up and three listening tests, lasted 45 min.

Subjects

Twenty-three volunteer listeners participated in this study. Each had been tested in Experiment 1, and each was a phonetic listener as designated by the identification test. All subjects were paid for their time. They were tested in groups of 2 to 6 at a time.

Results and Discussion

Four analyses of variance were performed to assess the effects of the spectrum conditions—whether identification varied across three-tone, single-tone, and single sawtooth presentations—for each of the target vowels. The mean proportions of the responses across the five spectrum conditions are shown in Figures 21, 22, 23, and 24. To summarize the outcomes, the identification of the three-tone targets differed from the identification of the single-component tones, and again from the component tones realized as sawtooth waves. In several cases, the identification of a single sinusoid differed from the matched single sawtooth.

The analysis in the case of the *bit* target revealed that the interaction of the factors Spectrum (three-tone, Tone 1, Tone 2, Sawtooth 1, Sawtooth 2) and Errors (*bat, bet, but:* the unintended response alternatives) was highly significant, $F(8, 176) = 11.87$, $p < .0001$. Figure 21 presents this condition. The three-tone target differed significantly from each of the component conditions, determined by post hoc comparisons of the error means (Newman-Keuls). Both Tone 1 and Sawtooth 1 elicited more identifications as UH; Tone 2 drew more identifications as AE, and Sawtooth 2 more identifications as EH. This shows clearly that the identification of the *bit* target is not simply attributable to the vowel likenesses of its component tones, though both versions of the second component seem overall like the vowel IH despite the increase in unintended responses. There were no differences observed between Tone 1 and Sawtooth 1 or between Tone 2 and Sawtooth 2, indicating that the

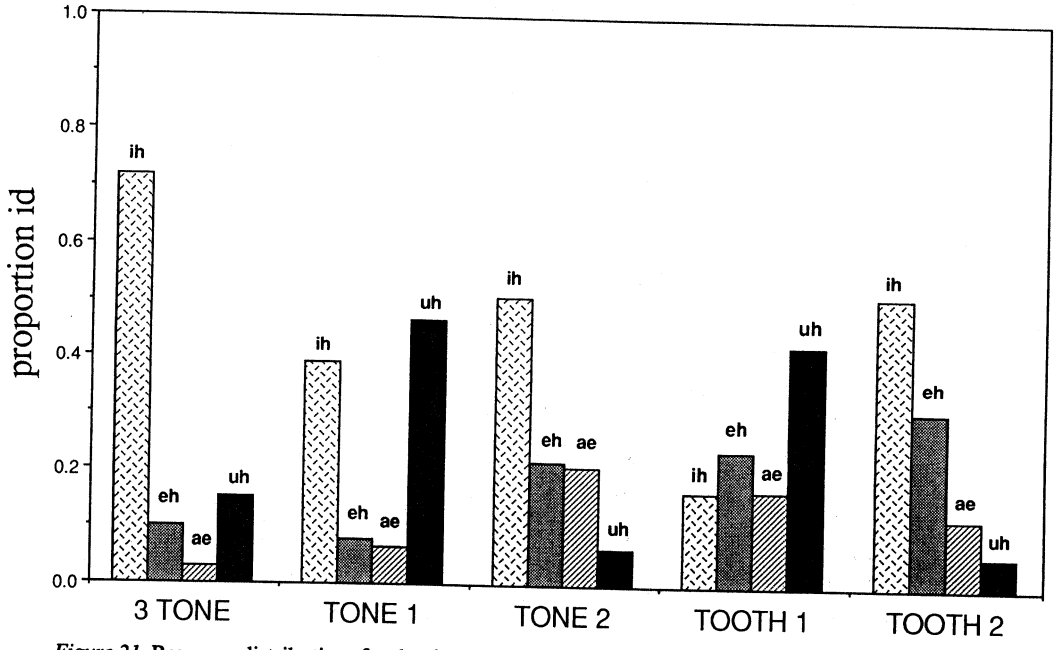

Figure 21. Response distributions for the sinusoidal targets realized as 3-tone patterns, Tone 1 and Tone 2 presented in isolation, and the two lowest tones synthesized as frequency modulated sawtooth waves, Sawtooth 1 and Sawtooth 2: This figure shows the results of the test with the *bit* target. (Proportion id = proportion of identification judgments.)

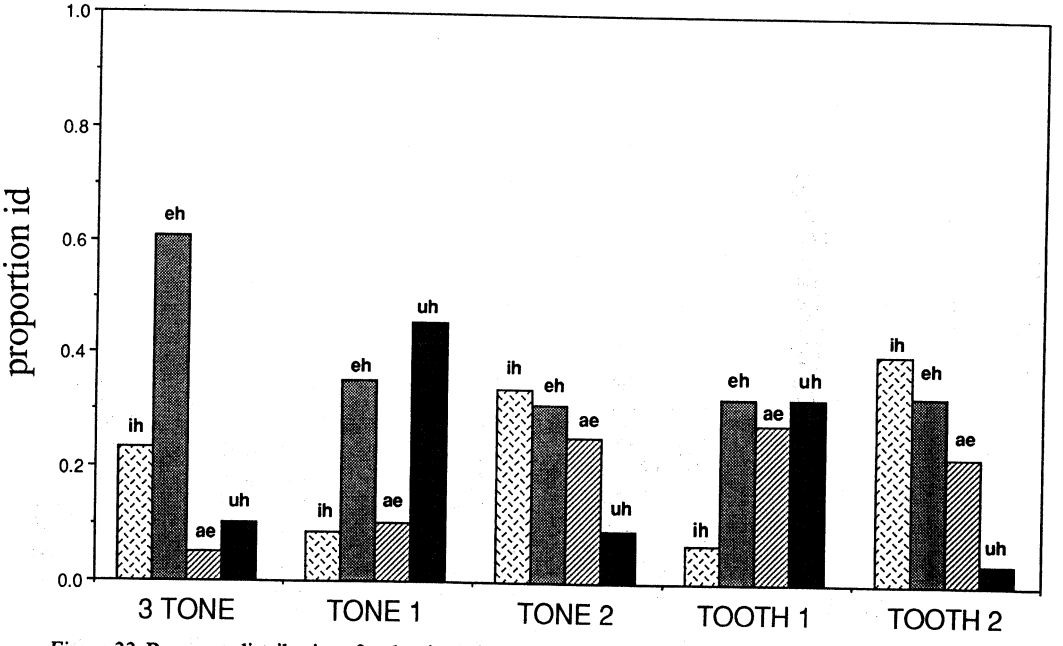

Figure 22. Response distributions for the sinusoidal targets realized as 3-tone patterns, Tone 1 and Tone 2 presented in isolation, and the two lowest tones synthesized as frequency modulated sawtooth waves, Sawtooth 1 and Sawtooth 2: This figure shows the results of the test with the *bet* target. (Proportion id = proportion of identification judgments.)

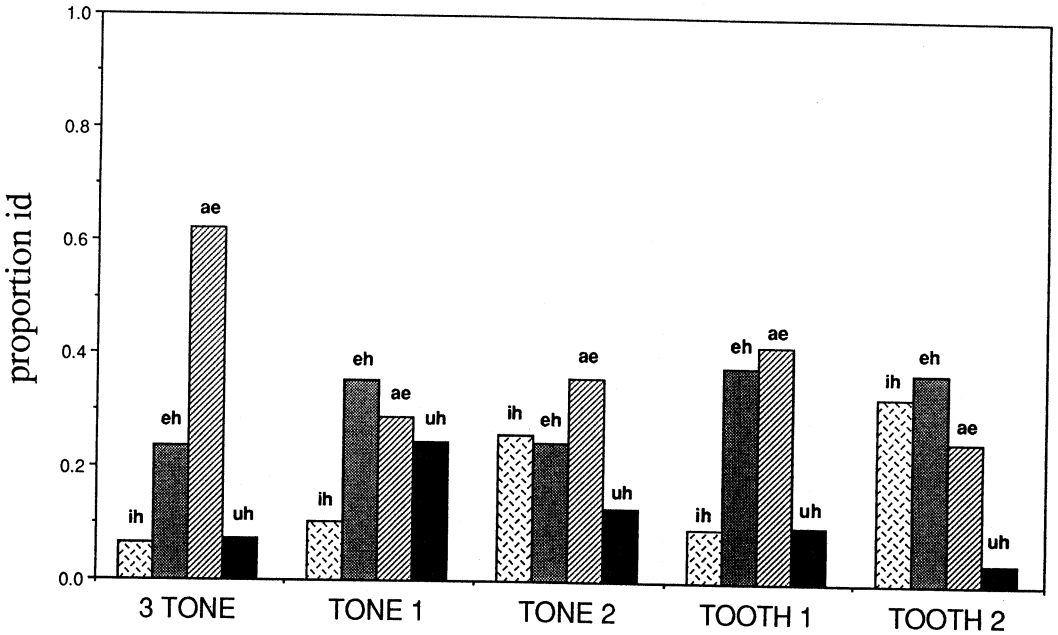

Figure 23. Response distributions for the sinusoidal targets realized as 3-tone patterns, Tone 1 and Tone 2 presented in isolation, and the two lowest tones synthesized as frequency modulated sawtooth waves, Sawtooth 1 and Sawtooth 2: This figure shows the results of the test with the *bat* target. (Proportion id = proportion of identification judgments.)

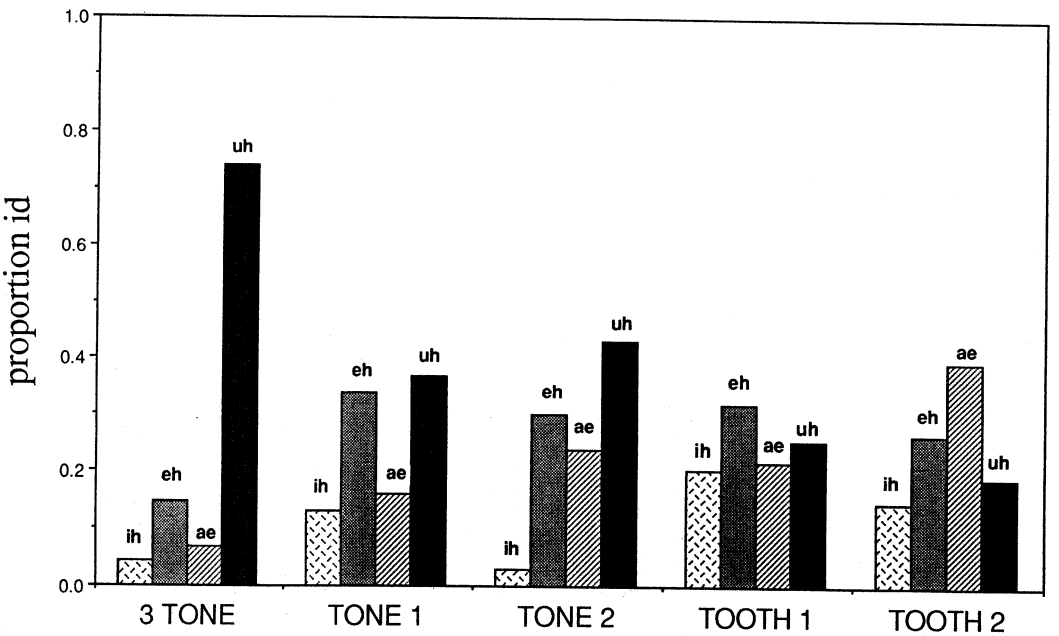

Figure 24. Response distributions for the sinusoidal targets realized as 3-tone patterns, Tone 1 and Tone 2 presented in isolation, and the two lowest tones synthesized as frequency modulated sawtooth waves, Sawtooth 1 and Sawtooth 2: This figure shows the results of the test with the *but* target. (Proportion id = proportion of identification judgments.)

frequency of the amplitude peak in the spectrum, rather than the shape of the spectrum envelope, contributed to the judgment.

The results in the case of the *bet* target are shown in Figure 22. Again, the Spectrum × Errors interaction was significant, $F(8, 176) = 11.64$, $p < .0001$. The identification of the three-tone pattern differed from each of the single-tone versions, with significantly more AE responses to Tone 2, Sawtooth 1, and Sawtooth 2; in the remaining comparison, Tone 1 was identified as UH quite often, also differing from the three-tone pattern. In this case, the imposition of the sawtooth wave on the pattern of Tone 1 proved to be perceptually distinctive, leading to more AE responses and fewer UH responses to Sawtooth 1 than to its mate, Tone 1. By implication, listeners in this case took the spectrum envelope into account in designating a vowel impression for the lowest tone presented in isolation.

The results in the case of the *bat* target are parallel to *bit* and *bet:* the Spectrum × Errors interaction was significant, $F(8, 176) = 4.90$, $p < .0001$. The three-tone identifications differed from each of the single-tone conditions, as is shown in Figure 23. There were significantly more UH responses to Tone 1 and IH responses to Tone 2 and Sawtooth 2; Sawtooth 1 elicited more identifications as EH than did the three-tone pattern. The spectrum envelope also influenced the perception of the lowest tone, inasmuch as Tone 1 was reported as UH more frequently than was Sawtooth 1. These findings again indicate that the perception of the vowel in the three-tone *bat* target is not reducible to the effects of its component tones.

Finally, the *but* target results appear consistent with the others, the three-tone pattern differing from the single tones in each instance. Here, once again, the Spectrum × Errors interaction was significant, $F(8, 176) = 2.45$, $p < .02$. Tone 1 and Sawtooth 1 both produced more EH responses, and Tone 2 and Sawtooth 2 more AE responses. Moreover, Tone 2 and Sawtooth 2 differed in the extent to which they evoked AE labels, which reflects the perceptual influence of spectrum envelope on the perceptual value of the frequency of the amplitude peak.

Conclusion

Overall, this test established that the reports of vowel identity of the three-tone targets do not match the reports for either of the two perceptually important components, Tone 1 and Tone 2, presented in isolation. Though the variation across the target conditions of the responses to Tone 2 approximates the three-tone case, the information that Tone 2 conveys about the vowel was much more effective in the acoustic context of Tone 1. Were this outcome to involve synthetic or edited natural speech signals, it would not be remarkable (Christovich, Sheikin, & Lublinskaja, 1979; Delattre, Liberman, Cooper, & Gerstman, 1952; but, see Klatt, 1982, 1985). But in view of the durability of the distinctly tonal qualities when three-tone complexes are perceived phonetically, this result provides a key to understanding the perceptual organization of our present target syllables. The attribution of vowel quality to the sinusoidal signals appears to use information simultaneously available from the two lowest tonal components. This finding distinguishes sinusoidal vowel perception from the attribution of speechlike qualities to spectrally simple nonspeech sounds (for a variant pertaining to perceptual learning, see Grunke & Pisoni, 1982). Although sinusoidal vowel perception appears to rely on acoustic information analogous to that which presumably occurs in ordinary speech signals, the imposition of vowel names on simple tones does not similarly depend on phonetic information. Assigning a vowel name to a simple nonspeech tone may truly require an assertion of remote resemblance between nonspeech and speech impressions, a perceptual circumstance distinct from the apprehension of phonetic information from an acoustic signal.

The comparison of sinusoid and sawtooth versions of the same frequency patterns produced both clear and puzzling results. On the one hand, subjects evidently are influenced by timbre when judging single tones, inasmuch as the frequency at which greatest power in the spectrum occurred was the same in matched sinusoidal and sawtooth instances. This result, noted in two cases involving the second tone and in one involving the first, had no consistent effect. Imposing a sawtooth waveform on the frequency pattern did not create the impression consistently of a lower or a higher vowel; nor did it consistently affect the vowel dimension of advancement; nor did it make the resulting vowel likeness seem consistently more central or less central. In the case of Tone 1 of *bat,* the imposition of the sawtooth waveform was tantamount to decreasing the frequency of the syllable nucleus; in the cases of Tone 1 of *bet* and Tone 2 of *but,* it was equivalent to increasing the frequency. The lack of a clear pattern to the results of this experimental manipulation requires no more than a speculation: Might the effect be specific to the task of judging single tones? Informally, we have observed that imposing sawtooth or triangle waveforms on three-tone patterns affects apparent naturalness, though not the intelligibility of the sentence. Whether the properties of the spectrum envelope create phonetic categorization effects more generally and whether the present outcome requires a phonetic or an auditory motivation remain, therefore, topics for further study.

General Discussion

Which properties of the acoustic speech signal provide information to the listener about the talker's message? Our studies of sinusoidal replicas of natural utterances offer a contrast to the perspective that the perceiver is a meticulous listener whose attention is devoted to the elemental attributes of speech signals. When these elements, which have often been presumed to be essential for perception, are removed from the signal and replaced by sinusoids, phonetic perception persists. Or so it seemed. The present experiments tested this proposition, that the listener transcribed sinusoidal sentences by virtue of ordinary perception, and in essence did not use an explicit strategy or rationalization. Because the perception of sinusoidal vowels in [bVt] syllables resembles vowel perception and because vowel identification of sinewave syllables involves the concurrent properties of the components of multitone patterns, it seems as though interconsonantal vowel perception with three-tone sinewave syllables is phonetically motivated. Moreover, because the perception of sinusoidal vowels was affected by the range of tone variation in precursor sentences, listeners may be said to organize three-tone replicas as if the signals were produced vocally. The frequency range effect is, in fact, a vocal-

tract scale effect, though certainly the "vocal tract" that issues sinusoidal speech is an abstract one, indeed.

The factors that regulate the perceptual apprehension of phonetic information from time-varying patterns remain to be specified. A full account will also require (a) an exposition of the perceptual organization of the component tones as a single coherent stream and (b) a means of relating these rather special listening conditions to the perceptual organization of natural speech. In the meanwhile, the finding that speech perception can endure the absence of short-time acoustic elements characteristic of vocal sound production places definite constraints on models of speech perception.

References

Ainsworth, W. A. (1974). The influence of precursive sequences on the perception of synthesized vowels. *Language and Speech, 17,* 103–109.

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

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.

Boring, E. G., (1942). *Sensation and perception in the history of experimental psychology.* New York: Appleton-Century.

Bricker, P. D., & Pruzansky, S. (1976). Speaker recognition. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 295–326). New York: Academic Press.

Chistovich, L. A., Sheikin, R. L., & Lublinskaja, V. V. (1979). "Centres of gravity" and spectral peaks as the determinants of vowel quality. In B. Lindblom & S. Ohman (Eds.), *Frontiers of speech communication research* (pp. 143–157). New York: Academic Press.

Cutting, J. E. (1974). Two left hemisphere mechanisms in speech perception. *Perception & Psychophysics, 16,* 601–612.

Darwin, C. J., & Baddeley, A. D. (1974). Acoustic memory and the perception of speech. *Cognitive Psychology, 6,* 41–60.

Dechovitz, D. R. (1977a). Information conveyed by vowels: A confirmation. *Haskins Laboratories Status Report on Speech Research, SR-51/52,* 213–219.

Dechovitz, D. R., (1977b). Information conveyed by vowels: A negative finding. *Journal of the Acoustical Society of America, 61,* S39.

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.

Delattre, P. C., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color. *Word, 8,* 195–210.

Disner, S. F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America, 67,* 253–261.

Eimas, P. D., & Miller, J. L. (1978). Effects of selective adaptation on the perception of speech and visual patterns: Evidence for feature detectors. In R. D. Walk & H. L. Pick, Jr. (Eds.), *Perception and experience* (pp. 307–345). New York: Plenum.

Elman, J. L., & McClelland, J. L. (1985). Exploiting lawful variability in the speech waveform. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360–385). Hillsdale, NJ: Erlbaum.

Fant, C. G. M. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics, 1,* 3–108.

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

Fant, C. G. M. (1966). A note on vocal tract size factors and nonuniform *F*-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report 4/66* (pp. 22–30). Stockholm, Sweden: Royal Institute of Technology. (Reprinted in 1973 as chapter 4 in *Speech sounds and features* (pp. 84–93). Cambridge, MA: MIT Press.)

Fourcin, A. J. (1968). Speech-source inference. *IEEE Transactions on Audio and Electroacoustics, ACC-16,* 65–67.

Gatewood, E. L. (1920). The vocality of fork, violin and piano tones. *American Journal of Psychology, 31,* 194–203.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics, ACC-16,* 78–80.

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

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

Joos, M. (1948). Acoustic phonetics. *Language, 24* (Supplement), 1–137.

Klatt, D. H. (1979). Speech perception: A model of acoustic–phonetic analysis and lexical access. *Jounal of Phonetics, 7,* 279–312.

Klatt, D. H. (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step. *Proceedings of the ICASSP, Paris,* 1278–1281.

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.

Köhler, W. (1910). Akustische Untersuchungen, II [Acoustical investigations]. *Zeitschrift für Psychologie, 58,* 59–140.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America, 66,* 1668–1679.

Ladefoged, P. (1967). The nature of vowel quality. In *Three areas of experimental phonetics* (pp. 50–142). London: Oxford University Press.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language, 56,* 485–502.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29,* 98–104.

Liberman, A. M. (1970). The grammars of speech and language. *Cognitive Psychology, 1,* 301–323.

Lieberman, P. (1984). *The biology and evolution of language.* Cambridge, MA: Harvard University Press.

Markel, J. D., & Gray, A. H., Jr. (1976). *Linear prediction of speech.* New York: Springer-Verlag.

Massaro, D. W. (1972). Perceptual images, processing time, and perceptual units in auditory perception. *Psychological Review, 79,* 124–145.

Modell, J. D., & Rich, G. J. (1915). A preliminary study of vowel qualities. *American Journal of Psychology, 26,* 453–456.

Nearey, T. M. (1978). *Phonetic feature systems for vowels.* Bloomington, IN: Indiana University Linguistics Club.

Papçun, G. (1980). *How do different speakers say the same vowels? Discriminant analyses of four imitation dialects.* UCLA Working Papers in Phonetics. Los Angeles: UCLA.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175–184.

Rand, T. C. (1971). Vocal tract size normalization in the perception of stop consonants. *Haskins Laboratories Status Reports on Speech Research, SR-25/26,* 141–146.

Remez, R. E. (in press). Units of organization and analysis in the per-

ception of speech. In M. E. H. Schouten (Ed.), *Psychophysics of speech perception.* Amsterdam: Martinius Nijhoff.

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., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212,* 947–950.

Rubin, P. E. (1971). *Psychological space of vowel perception.* Unpublished senior thesis, Brandeis University, Waltham, MA.

Rubin, P. E. (1980). *Sinewave synthesis.* Internal memorandum, Haskins Laboratories, New Haven, CT.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics, 31,* 307–314.

Shankweiler, D., Strange, W., & Verbrugge, R. R. (1977). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing* (pp. 315–345). Hillsdale, NJ: Erlbaum.

Shepard, R. N. (1972). Psychological representation of speech sounds. In E. E. David, Jr., & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 67–113). New York: McGraw-Hill.

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.

Summerfield, A. Q., & Haggard, M. P. (1973). Vocal tract normalization as demonstrated by reaction times. *Speech Perception: Report on Work in Progress* (Queen's University of Belfast), *2*(3), 1–26.

Suomi, K. (1984). On talker and phoneme information conveyed by vowels: A whole spectrum approach to the normalization problem. *Speech Communication, 3,* 199–209.

Sussman, H. M. (1986). A neuronal model of vowel normalization and representation. *Brain and Language, 28,* 12–33.

Syrdal, A. K., & Gopal, H. S. (1985). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79,* 1086–1100.

Thompson, C. L., & Hollien, H. (1970). Some contextual effects on the perception of synthetic vowels. *Language and speech, 13,* 1–13.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America, 60,* 198–212.

Wakita, H. (1977). Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-25,* 183–192.

Williams, D. R., Verbrugge, R. R., & Studdert-Kennedy, M. (1983). Judging sine wave stimuli as speech and as nonspeech. *Journal of the Acoustical Society of America, 74,* S36.

Zue, V. W., & Schwartz, R. M. (1980). Acoustic processing and phonetic analysis. In W. A. Lea (Ed.), *Trends in speech recognition* (pp. 101–124). Englewood Cliffs, NJ: Prentice-Hall.

Received March 20, 1986
Revision received September 3, 1986 ■