Research Article

ON THE BISTABILITY OF SINE WAVE ANALOGUES OF SPEECH

Robert E. Remez,[1] Jennifer S. Pardo,[2] Rebecca L. Piorkowski,[1] and Philip E. Rubin[3]

[1]Barnard College, [2]Yale University, and [3]Haskins Laboratories

Abstract—*Our studies revealed two stable modes of perceptual organization, one based on attributes of auditory sensory elements and another based on attributes of patterned sensory variation composed by the aggregation of sensory elements. In a dual-task method, listeners attended concurrently to both aspects, component and pattern, of a sine wave analogue of a word. Organization of elements was indexed by several single-mode tests of auditory form perception to verify the perceptual segregation of either an individual formant of a synthetic word or a tonal component of a sinusoidal word analogue. Organization of patterned variation was indexed by a test of lexical identification. The results show the independence of the perception of auditory and phonetic form, which appear to be differently organized concurrent effects of the same acoustic cause.*

How does a listener know what a talker just said? A fundamental perceptual component in acts of spoken communication is the analysis of sensory samples of speech. However, perception of the phonetic properties in stimulation cannot proceed as if sensory activity stems from speech sources alone. People speak and listen to each other amid multiple sources of sound. Indeed, the vocal apparatus itself is a source of respiratory and ingestive sound as well as speech. In this respect, the perception of speech naturally entails two functions: (a) an organizational function that identifies a sensory pattern attributable to a spoken source and (b) an analytical function that identifies the phonetic attributes conveyed in a sensory pattern.

Although there are many descriptions of phonetic analysis (catalogued in Klatt, 1989), the organizational component of perception is typically described in a general *auditory* account based on similarity and continuity grouping (Bregman, 1990; Darwin, 1997). In this conceptualization, an automatic process parses simultaneous and successive sensory elements into groups of like elements, and evidence of its action derives from a listener's perception of simple test patterns of tones and noise. The adequacy of this account is doubtful for speech signals, because it falsely presumes that the acoustic elements of a single speech stream exhibit mutual likeness. The whistles, clicks, hisses, buzzes, and hums that compose a speech stream are arguably grouped perceptually by the action of an alternative *phonetic* organizational principle keyed to complex spectrotemporal patterns despite dissimilar elementary constituents (Remez, Rubin, Berns, Pardo, & Lang, 1994). Although our findings have encouraged such a formal account, we lack direct perceptual evidence of two alternative principles of coherence, one that organizes patterns by similarity and another that organizes patterns by orderly albeit complex variation.

One difficulty in assessing the independence of auditory and phonetic organization arises because the two modes can converge, thereby concealing their distinct action. Indirect evidence of alternative organizations can be found in reports about duplex perception of synthetic speech, in which a spatially isolated acoustic element is organized in two ways: (a) segregated into a perceptual stream, yielding an impression of its auditory form, and (b) fused with concurrent acoustic components into a speech stream, yielding impressions of phonetic identity (Rand, 1974). However, few tests have sought to determine whether the two modes of organization occur concurrently, and therefore not contingently (Liberman, Isenberg, & Rakerd, 1981; Whalen & Liberman, 1987, 1996; Xu, Liberman, & Whalen, 1997). To provide direct evidence of two concurrent alternative modes of perceptual organization, our tests used sine wave analogues of speech, in which auditory impressions of acoustic constituents persist during phonetic organization.

Nonphonetic auditory impressions of sine wave analogues are correlated with the acoustic features of these signals—in the absence of phonetic perception, sine wave analogues sound like several simultaneous whistles (Remez, Rubin, Pisoni, & Carrell, 1981). Few listeners notice phonetic attributes unless they are asked specifically to transcribe computer-generated speech. When listeners do recognize the message in a sine wave sentence, this phonetic organization accompanies reports that the vocal quality remains highly unnatural. In this report, we offer evidence that auditory impressions of sine wave analogues are independent of phonetic organization by assessing the hypothetical bistability of sine wave analogues of speech.

The present study consisted of two experiments. The first examined the assertion that synthetic speech is organized in a single stable form that precludes the resolution of its vocalic constituents, the formants. One condition tested a listener's resolution of the frequency variation of an isolated second formant using a simple same/different task. A second condition tested perceptual resolution of this acoustic element when it occurred within a synthetic word. In the second experiment, we determined whether a listener could resolve the auditory form of a single tone constituent of a sinusoidally replicated word in two conditions. The first assessed sine wave tone resolution without a listener's awareness of the phonetic properties of tone ensembles, thereby permitting a measure of auditory form perception in the absence of phonetic organization (see Remez et al., 1981). In the second condition, performance on the auditory form task was assessed in a dual-task procedure, which included both tone and word verification.

Listeners in the tests using synthetic speech were able to match an individual formant pattern when it was presented in isolation, but not when it occurred within a phonetically effective acoustic pattern. Listeners in the tests using time-varying sinusoids, however, resolved the auditory form of a component tone despite concurrent phonetic fusion. The results of these studies establish that auditory organization and phonetic organization of synthetic speech coincide, in contrast to the discrepant auditory and phonetic organization of perceptually bistable sine wave analogues.

Address correspondence to Robert E. Remez, Department of Psychology, Barnard College, 3009 Broadway, New York, NY 10027-6598; e-mail: remez@columbia.edu.

EXPERIMENT 1: SYNTHETIC SPEECH

Method

Materials

Nine words (*beak*, *sill*, *wed*, *pass*, *lark*, *rust*, *jaw*, *shook*, and *coop*), each with a different vowel, composed the test set. The words were spoken by one of the authors (R.E.R.), recorded on audiotape, digitized, and analyzed to create synthesis parameters for the KLSYN88a speech synthesizer (implemented by Sensimetrics, Inc., as SenSyn PPC).

To guarantee that each synthetic word would exhibit the same vocal pitch contour, we first analyzed the fundamental frequency of the natural utterance of *shook*. Taking this pattern as a model, we created a standard pattern for each of the synthetic words, dividing the 350 ms of voicing in *shook* into five equal spans, estimating the onset and offset frequencies, and, scaling the pattern to each of the synthetic words. SenSyn interpolated the frequencies between the junctures.

Nine voiced second-formant patterns were synthesized in isolation from their phonetic patterns, for use in a test of perceptual resolution of formant patterns. Isolated second-formant patterns did not elicit phonetic impressions, and exhibited a buzzing timbre.[1]

Procedure

We used two tasks to examine perceptual organization of synthetic speech. First, we assessed baseline auditory resolution of the frequency variation of isolated second-formant patterns. On each trial, a listener heard two isolated second-formant patterns separated by 300 ms and indicated whether their auditory form was the same or different, as shown in Figure 1a. On *same* trials, the second-formant patterns were identical. On *different* trials, the second-formant patterns were derived from synthetic words that differed by a single vowel step, defined by ordering the vowels by the average frequency of the second formant: /i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /ʌ/, /ɔ/, /ʊ/, and /u/.

The second task assessed organization of a second-formant pattern within a synthetic speech signal. On each trial, a listener heard an isolated second-formant pattern followed by a synthetic word and reported whether the pattern was a component of the synthetic word, as shown in Figure 1b. On *yes* trials, the isolated second-formant pattern was part of the synthetic word. On *no* trials, the isolated second-formant pattern did not occur within the word, which differed from the source of the isolated second formant by one vowel step, as in the first task. Listeners completed both tasks (144 trials each) in a single session and indicated their impressions by marking an answer booklet. They heard the materials presented through headphones via digital audiotape.

Subjects

Twenty-four listeners from the undergraduate population of Columbia University received credit toward a course requirement for participating. All were native speakers of English and reported normal hearing at the time of testing. None had participated in experiments using synthetic speech. Five listeners were excluded from consideration because they failed to answer on all trials.

1. Examples of the acoustic test material used in this study are available for listening on the World Wide Web at http://www.columbia.edu/~remez/bistability.html.

Results and Discussion

The performance of each listener was represented as a value of d' for each test item (Macmillan & Creelman, 1991) in each task. Figure 2a shows the group mean d' scores for tests of formant pattern resolution in isolation and within a synthetic word. Listeners were able to resolve the auditory form of a synthetic second-formant pattern presented in isolation, as exhibited by a high mean d' score. In contrast, listeners could not resolve the same formant pattern embedded in a synthetic word, with a mean d' score approaching 0. The difference in performance on the two tasks was verified by an analysis of variance (one-way) on d', indicating a main effect for task, $F(1, 16) = 171.13$, $p < .0004$.

This study provides evidence that listeners are unable to verify the auditory form of a second formant when listening to synthetic speech. This finding cannot be ascribed to insensitivity to frequency variation of the second formant, because the listeners were able to resolve the isolated second-formant patterns. Moreover, researchers have deduced that listeners must be sensitive to this property of the spectro-temporal pattern of speech by virtue of the correlation of frequency variation in the second formant with consonantal place of articulation (Cooper, Delattre, Liberman, Borst, & Gerstman, 1952) and vowel advancement (Peterson & Barney, 1952). Synthetic speech evidently evokes auditory perceptual organization in which the acoustic elements, despite their discontinuity and dissimilarity, are bound together.

To study the divergent effects of perceptual organization, one cannot rely on synthetic speech alone, because the auditory constituents of these signals cohere. Julesz and Hirsh (1972) discussed this aspect of the speech signal:

> The harmonics of the voice are equally separated on a linear scale of frequency, but certain groups of them get reinforced by resonant properties of the vocal tract. These 'formants' do not stand out as separate figures but turn out instead to be the bases for identifying the spoken vowels. . . . Whether the structural features in spoken sound patterns show this [perceptual coherence] by virtue of properties of the stimulus configuration or of the language habits of the listener is not clear. (pp. 300–305)

Although the perceptual coherence noted by Julesz and Hirsh was demonstrated for the elements of synthetic speech in Experiment 1, sine wave analogues of speech may differ in their organization; the auditory form of tonal constituents might be resolved while the cohesion requisite for phonetic perception occurs. In Experiment 2, we tested this hypothetical bistability of sine wave analogues of speech.

EXPERIMENT 2: SINE WAVE ANALOGUES OF SPEECH

Method

Materials

Nine sine wave words were modeled on the same natural tokens used to make the synthetic speech in Experiment 1, but the acoustic analysis was used to compose synthesis parameters for a sine wave synthesizer (Rubin, 1980). The patterns used one sine wave for the frequency and amplitude of each of the three lowest-frequency formants in each of the words; in the four words that included a fricative consonant (*sill*, *pass*, *rust*, and *shook*), a fourth sine wave was used to replicate the frequency and amplitude pattern of the fricative formant.

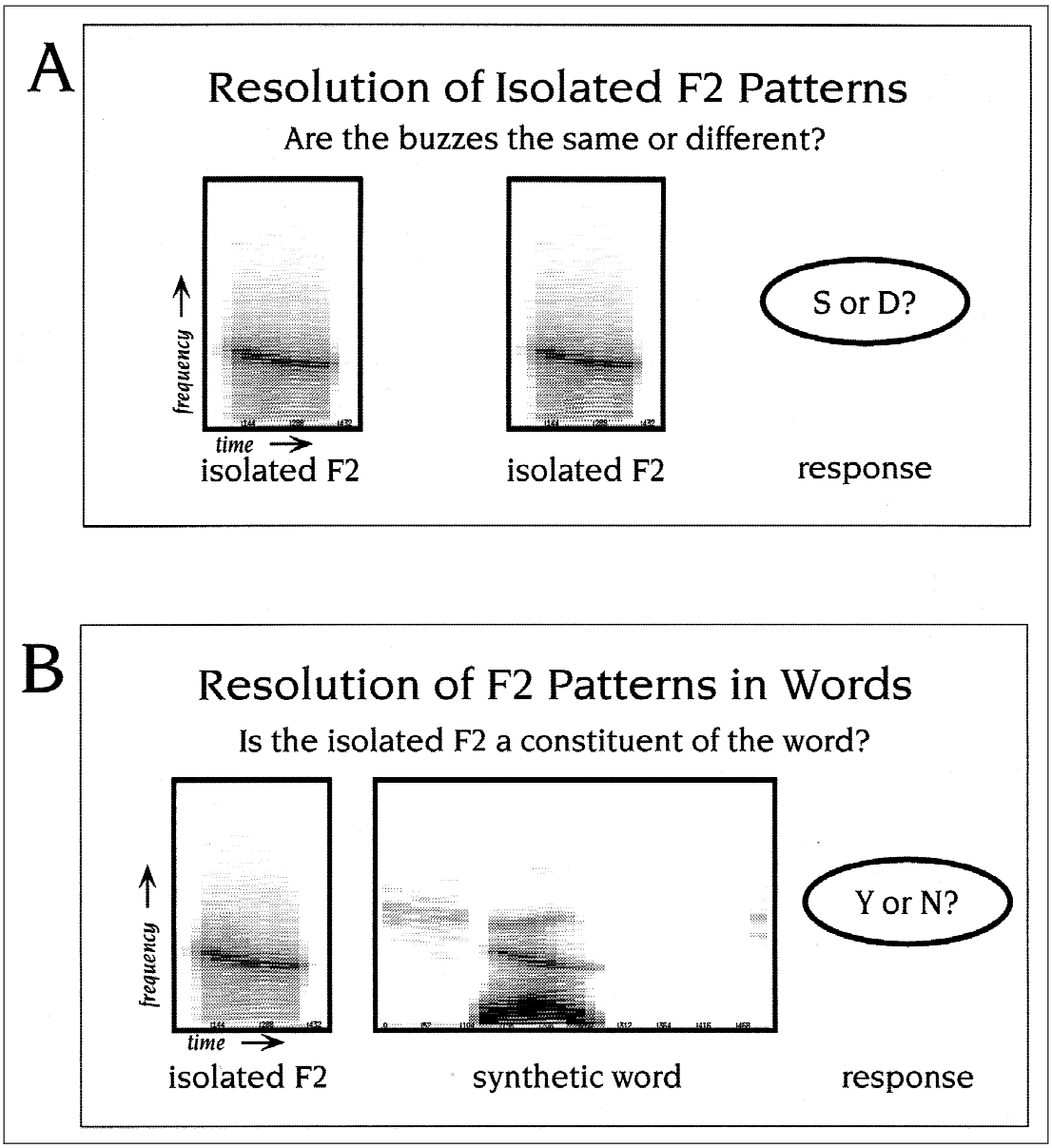

Fig. 1. Trial formats from Experiment 1. In the test of resolution of isolated second-formant (F2) patterns (a), a subject indicated whether the two buzzes were the "same" (S) or "different" (D). In the test of resolution of second-formant patterns in words (b), a subject responded "yes" (Y) or "no" (N) to indicate whether the isolated formant occurred in the word that followed.

(A more complete description of sine wave replication of natural speech is provided in Remez, Rubin, Nygaard, & Howell, 1987.) Paralleling the test items in Experiment 1, the single tone patterns used as probes for the tone verification task in Experiment 2 consisted of the nine tonal analogues of the second formant.

Procedure

We designed two tasks to assess auditory and phonetic perceptual organization over the course of four testing sessions. A tone verification task (Sessions 1 and 2) was used to measure the resolution of auditory form without phonetic perception: A sample tone preceded the presentation of the tone complex by 300 ms, and a listener indicated whether or not the sample tone was a component of the tone complex, as shown in Figure 3a. On *yes* trials, the sample tone was the second-formant analogue of the ensemble replicating a word. On *no* trials, the sample tone was the second-formant analogue of a different sinusoidal word in the test set—specifically, the second-formant analogue of the word that differed from the target by a single vowel step, defined by ordering the vowels by the average frequency of the second formant.

In the dual tone and word verification task (Sessions 3 and 4), a listener indicated whether a sample tone occurred in the complex of tones composing a word and also whether the sinusoidal word replica was the same as a printed word, as shown in Figure 3b. Much as in the

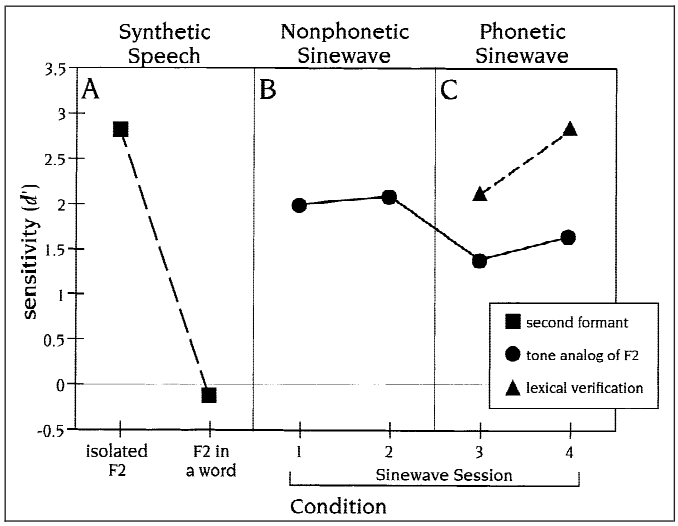

Fig. 2. Group performance in Experiments 1 (a) and 2 (b and c). Experiment 1 tested resolution of the second formant (F2) in isolation and in a word. Experiment 2 tested resolution of the tone analogue of the second formant within a tone complex not perceived to exhibit phonetic properties (b) and within a tone complex perceived phonetically (c); concurrent performance in a test of lexical verification of sine wave words is also shown.

tone verification task, the sample word either was identical to the sinusoidal word or differed from it by one step along the vowel series. Neither the tone verification task nor the word verification task was considered primary, and listeners were instructed to mark the responses in any order. Each of the testing sessions consisted of 144 trials and occurred on a separate day; no more than 2 days separated consecutive sessions for an individual subject.

Subjects

Twenty-two volunteer listeners were enlisted from the student population of Columbia University and were paid for their participation. All were native speakers of English and reported normal hearing at the time of testing. None had participated in studies using sinusoidal utterance replicas. The data from 3 subjects were excluded because they failed to complete all testing sessions.

Results and Discussion

Performance of each subject in each session was represented as a value of d' for each of the nine targets (Macmillan & Creelman, 1991). Figures 2b and 2c display group mean d' scores for tone verification performance over the four testing sessions. The high d' values indicate that listeners were able to resolve the pattern of a tonal analogue of the second formant, both when listening to a sine wave signal as a nonphonetic auditory form (Sessions 1 and 2) and when simultaneously perceiving the phonetic form of a word (Sessions 3 and 4). Indeed, performance on the tone verification task remained at a high level when the word verification task was concurrent. This was confirmed by a one-way analysis of variance (on the factor test block) on d' for tone, $F(3, 54) = 18.19$, $p < .001$, and a post hoc means test indicating that in all conditions performance was significantly different from chance ($p < .001$, Scheffé).[2]

The primary conclusion warranted by these data is that performance on both the tone and word verification tasks is very good, and that two kinds of organization, auditory and phonetic, occur simultaneously with sinusoidal analogues of words. While the sine wave components segregate into perceptual streams, sustaining performance on the tone verification task, they also fuse, promoting phonetic perception. This organizational divergence indicates the separate nature of early phonetic and auditory organization in perception (see also Remez et al., 1994). Such evidence warrants a conclusion that the perception of an acoustic pattern formed by a sine wave analogue is bistable, and differs from the organization of synthetic speech.

GENERAL DISCUSSION

Auditory Organization

These results confirm the difference in perceptual organization of sine wave analogues and synthetic speech. This difference is possible because some acoustic elements of synthetic speech cohere auditorily in the harmonically dense, amplitude-comodulated signals issuing from a vocal source (Assmann & Summerfield, 1990; Carrell & Opie, 1992). In contrast, the harmonically unrelated, asynchronously varying sine wave elements in replicated words do not cohere auditorily. Although sine wave analogues and synthetic speech signals differ with respect to their auditory form, each evokes phonetic perception.

In addition, these results show that the auditory organization of a sine wave analogue persists despite concurrent phonetic organization. In Experiment 2, listeners heard the pitch pattern of the tone analogue of a second formant while implicitly treating the same tone as information about place of articulation and vowel advancement. If impressions of auditory pitch stem from a different perceptual domain than phonetic impressions (Fodor, 1983), then the psychoacoustic evaluation of the auditory form of a sine wave analogue should yield the same function whether or not the pattern is also organized phonetically. Indeed, in tasks that direct attention to auditory qualities of speechlike tone patterns, perceptual sensitivity to acoustic properties remains consistent (see Johnson & Ralston, 1994; Kingston & Kirk, 1997; Sawusch & Gagnon, 1995; cf. Best, Studdert-Kennedy, Manuel, & Rubin-Spitz, 1989).

Accordingly, it is arguable that phonetic perception does not depend on the auditory form of an utterance. Whether a signal consists of a spectrotemporal pattern of sine wave tones or of harmonically related broadband resonances with interspersed episodes of noise, the perceiver treats it as a dynamically specified, phonologically governed act of a talker.

2. The data analysis also shows that performance on the tone verification task was impaired by the simultaneous word verification task. However, this decrement in tone verification is attributable to the concurrent execution of a linguistic task. In a control test not reported here, listeners were informed of the phonetic content of the tone ensembles, but performed only a tone verification task. Under these conditions, no performance decrement was observed. We report dual-task performance as evidence of simultaneous auditory and phonetic perceptual organization.

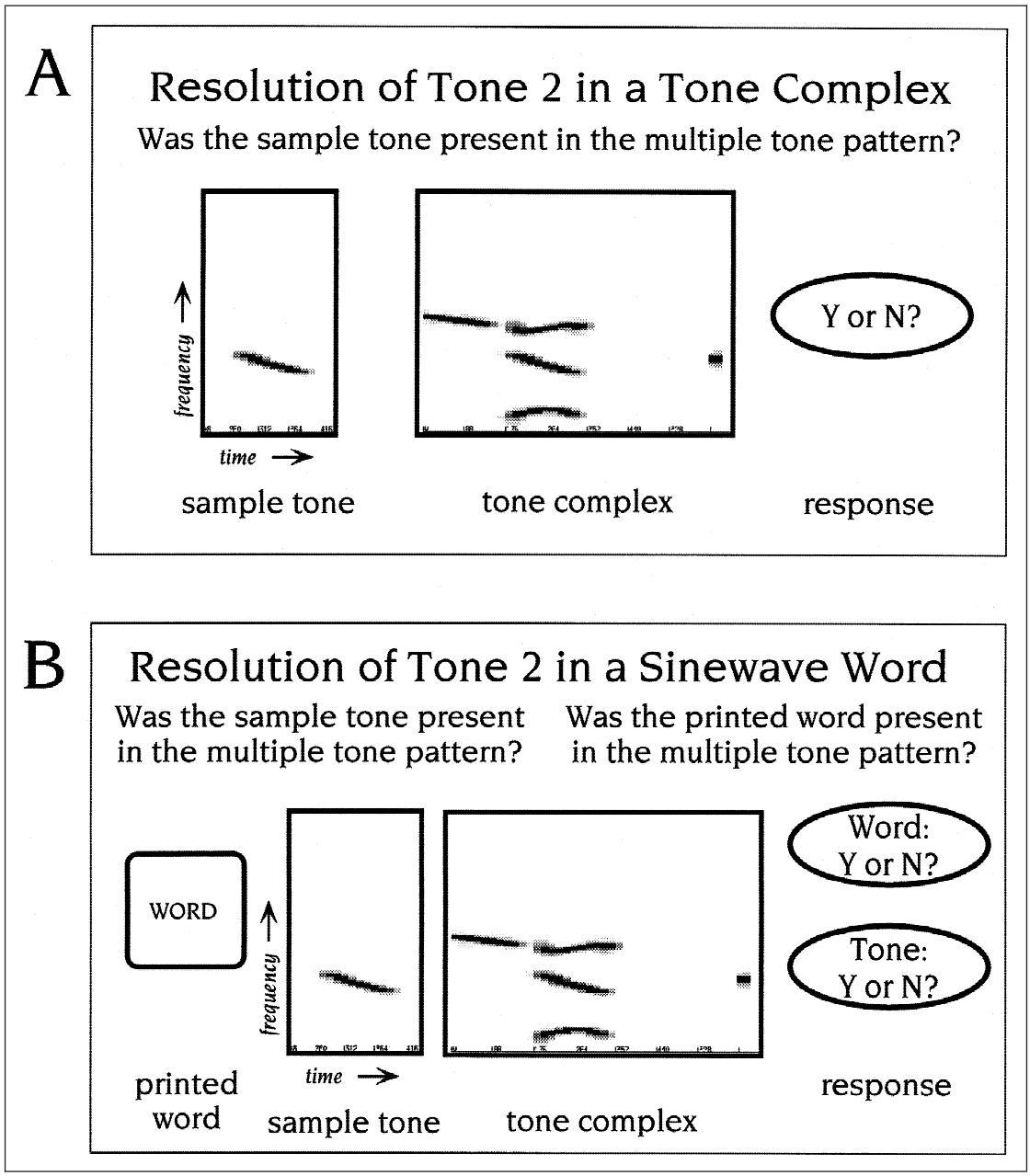

Fig. 3. Trial formats from Experiment 2. In the test of tone resolution in patterns perceived without phonetic attributes (a), a subject responded "yes" (Y) or "no" (N) to indicate whether the sample tone was present in the tone complex. In the concurrent tone resolution and lexical verification task (b), a subject responded "yes" or "no" to indicate if the sample tone was present in the tone complex and also to indicate if the printed word was present in the tone complex.

Phonetic Organization

Perceptual organization of sine wave analogues is an anomaly for accounts of phonetic perception that rely on auditory representation of the likely acoustic properties of natural speech (Diehl, Kluender, Walsh, & Parker, 1991). Sinusoidal replication preserves only coarse-grain spectrotemporal properties of speech signals while discarding the momentary acoustic constituents of utterances. The fact that a listener can transcribe a sine wave analogue reflects the use of dynamic information in phonetic perception.

Does the phonetic perceptual organization of sine wave analogues and natural speech occur through a common auditory process? This is proposed by Barker and Cooke (1997), whose automatic speech recognizer derives an all-pole representation of natural speech, which is then used to recognize sine wave analogues of speech, albeit more poorly than when the recognizer is trained on sine wave analogues.

However, the evidence reported here prescribes a different conclusion, that phonetic perception in humans does not rely on such a representation. If perceptual organization of sine wave analogues and of natural speech were governed solely by a common auditory process, then listeners would resolve the formants in natural speech much as they do the tones of sine wave analogues.

Furthermore, the phonetic effectiveness of the quantized noise-band signals of Shannon, Zeng, Kamath, Wygonski, and Ekelid (1995) also requires an alternative to a conventional auditory account of phonetic organization. In this technique, four noise bands span the range of 0 to 4 kHz, and each band is amplitude modulated at the level of the corresponding band in the natural spectrum of a speech sample. This technique removes most of the spectral detail of natural speech, leaving only amplitude variation in broad frequency bands. A spectral-peak-based recognizer (Barker & Cooke, 1997) trained on natural speech might fare poorly on a test using a signal with such stationary spectral bands, although human listeners readily comprehend such unnatural signals. The asynchronously varying amplitude-modulated noise bands composing this kind of signal frustrate the simple devices available in a Gestalt-based account of organization, as sine wave replicas of speech do (Remez et al., 1994).

In conclusion, our study revealed that a sine wave analogue of an utterance is perceptually bistable, unlike synthetic speech. Phonetic organization of sine wave analogues occurs independently of auditory organization, as informal evidence had suggested, and as these direct tests show. While showing that the perceptual organization of sine wave analogues is similar to the phenomenon of duplex perception, this study verified that divergent auditory and phonetic organizations are sustained simultaneously. The difference in auditory quality of natural speech and sine wave analogues tempers any proposal that intelligibility depends directly on auditory form. The high performance levels of untrained listeners is evidence that the perceptual functions called upon by our simple measures are used in ordinary perception.

Acknowledgments—For help and encouragement in developing this project, the authors thank Stefanie Berns, Robert Crowder, Jennifer Fellowes, Bella Schanzer, Sam Glucksberg, and the reviewers of this submission. This research was supported by a grant from the National Institute on Deafness and Other Communicative Disorders (DC00308, to Barnard College) and by a grant from the National Institute of Child Health and Human Development (HD01994, to Haskins Laboratories).

REFERENCES

Assmann, P.F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequency. *Journal of the Acoustical Society of America*, *88*, 680–697.

Barker, J., & Cooke, M. (1997, September). *Modeling the recognition of spectrally reduced speech*. Paper presented at the 1997 Eurospeech Conference in Rhodes, Greece.

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

Bregman, A.S. (1990). *Auditory scene analysis*. Cambridge, MA: MIT Press.

Carrell, T.D., & Opie, J.M. (1992). The effect of amplitude comodulation on auditory object formation in sentence perception. *Perception & Psychophysics*, *52*, 437–445.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., & Gerstman, L.J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*, *24*, 597–606.

Darwin, C.J. (1997). Auditory grouping. *Trends in Cognitive Science*, *1*, 327–333.

Diehl, R.L., Kluender, K.R., Walsh, M.A., & Parker, E.M. (1991). Auditory enhancement in speech perception and phonology. In R.R. Hoffman & D.S. Palermo (Eds.), *Cognition and the symbolic process: Analytical and ecological perspectives* (pp. 59–76). Hillsdale, NJ: Erlbaum.

Fodor, J.A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Johnson, K., & Ralston, J.V. (1994). Automaticity in speech perception: Some speech/nonspeech comparisons. *Phonetica, 51,* 195–209.

Julesz, B., & Hirsh, I.J. (1972). Visual and auditory perception—An essay of comparison. In E.E. David & P.B. Denes (Eds.), *Human communication: A unified view* (pp. 283–340). New York: McGraw-Hill.

Kingston, J., & Kirk, C.J. (1997). Must sine-wave analogues be perceived as implicit speech to be perceived categorically? *Journal of the Acoustical Society of America*, *102*, 3091.

Klatt, D.H. (1989). Review of selected models of speech perception. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 169–226). Cambridge, MA: MIT Press.

Liberman, A.M., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception & Psychophysics*, *30*, 133–143.

Macmillan, N.A., & Creelman, C.D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

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. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, *55*, 678–680.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., & Lang, J.M. (1994). On the perceptual organization of speech. *Psychological Review*, *101*, 129–156.

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

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

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

Sawusch, J.R., & Gagnon, D.A. (1995). Auditory coding, cues, and coherence in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 635–652.

Shannon, R.V., Zeng, F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.

Whalen, D.H., & Liberman, A.M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, *237*, 169–171.

Whalen, D.H., & Liberman, A.M. (1996). Limits on phonetic integration in duplex perception. *Perception & Psychophysics*, *58*, 857–870.

Xu, Y., Liberman, A.M., & Whalen, D.H. (1997). On the immediacy of phonetic perception. *Psychological Science*, *8*, 358–362.

(RECEIVED 5/10/99; REVISION ACCEPTED 5/2/00)