Perception of sinewave vowels

James M. Hillenbrand,[a] Michael J. Clark, and Carter A. Baer
Department of Speech Pathology and Audiology, Western Michigan University, Kalamazoo, Michigan 49008

(Received 7 September 2010; revised 10 March 2011; accepted 10 March 2011)

There is a significant body of research examining the intelligibility of sinusoidal replicas of natural speech. Discussion has followed about what the sinewave speech phenomenon might imply about the mechanisms underlying phonetic recognition. However, most of this work has been conducted using sentence material, making it unclear what the contributions are of listeners' use of linguistic constraints versus lower level phonetic mechanisms. This study was designed to measure vowel intelligibility using sinusoidal replicas of naturally spoken vowels. The sinusoidal signals were modeled after 300 /hVd/ syllables spoken by men, women, and children. Students enrolled in an introductory phonetics course served as listeners. Recognition rates for the sinusoidal vowels averaged 55%, which is much lower than the ~95% intelligibility of the original signals. Attempts to improve performance using three different training methods met with modest success, with post-training recognition rates rising by ~5–11 percentage points. Follow-up work showed that more extensive training produced further improvements, with performance leveling off at ~73%–74%. Finally, modeling work showed that a fairly simple pattern-matching algorithm trained on naturally spoken vowels classified sinewave vowels with 78.3% accuracy, showing that the sinewave speech phenomenon does not necessarily rule out template matching as a mechanism underlying phonetic recognition. © *2011 Acoustical Society of America*. [DOI: 10.1121/1.3573980]

PACS number(s): 43.71.Es, 43.71.An [PEI] Pages: 3991–4000

I. INTRODUCTION

It is well known that speech can remain intelligible in spite of signal manipulations that obscure or distort acoustic features that have been shown to convey critical phonetic information. One of the more striking demonstrations of this phenomenon comes from sinewave speech (SWS). In SWS a replica is made of an utterance by mixing sinusoids that follow the contours of the three or four lowest formant frequencies (Fig. 1). Although sinewave sentences sound quite strange, it is important to note that this synthesis approach has quite a bit in common with speech produced by the Pattern Playback (Cooper *et al.*, 1952) and with formant-synthesized speech (Klatt, 1980). In all three cases it is the formant frequency pattern that is the primary connection between the original and resynthesized utterances. The odd quality of SWS is due primarily to two major departures from Pattern Playback and formant-synthesized speech. First, in SWS the formant-simulating sinusoids are harmonically unrelated, resulting in an aperiodic signal. Natural speech, of course, consists of both quasiperiodic and aperiodic elements, along with segments such as voiced fricatives and breathy vowels which combine both periodic and aperiodic elements. Second, the simulated formant peaks of SWS are much narrower than the formants of either natural speech or speech reconstructed using either the Pattern Playback or a formant-synthesizer. In spite of their peculiar and quite unfamiliar sound quality, sinewave replicas of sentences are intelligible at some level. In their original study, Remez *et al.* (1981), using a single sinewave sentence ("Where were you a year ago?"), asked listeners for their spontaneous impressions of the stimulus with no special instructions about the nature of the signals they would be hearing. Nearly half of the listeners heard the stimuli as speech-like, describing it variously as human speech, human vocalizations, artificial speech, or reversed speech. Strikingly, two of the 18 subjects not only heard the stimulus as speech but also transcribed the sentence accurately. Intelligibility improved considerably when listeners were instructed to hear the signals as speech, although a substantial number of listeners were still unable to transcribe the sentence accurately, and still others did not hear it as speech even with instructions to do so. The intelligibility of sinewave sentences has since been tested in many experiments. Carrell and Opie (1992), for example, tested listeners on four simple sinewave sentences consisting entirely of sonorants (e.g., "A yellow lion roared."), with instructions to hear the stimuli as speech and transcribe as much of the utterance as possible. Intelligibility averaged about 60% for three of the four sentences and was nearly perfect for the remaining sentence.

In short, sinewave sentences are intelligible, though imperfectly so. To what should we attribute the intelligibility of these odd sounding signals and what does the SWS phenomenon say about the underlying pattern-matching mechanisms that are involved in speech recognition? SWS findings are frequently discussed in terms of what they might tell us about phonetic recognition (e.g., Remez *et al.*, 1981), but because most of this work has used meaningful and syntactically well-formed sentences, which are subject to well-known effects of higher level linguistic knowledge, it remains unclear how much expressly phonetic information

[a] Author to whom correspondence should be addressed. Electronic mail: james.hillenbrand@wmich.edu

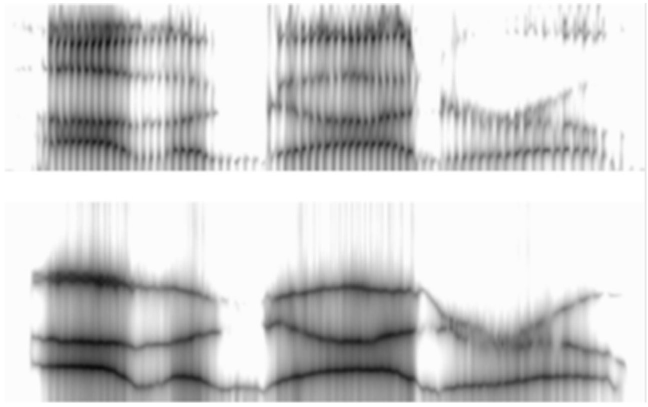

FIG. 1. Spectrogram of the naturally spoken utterance "Avagadro" (top) and a sinewave replica of the utterance.

the listeners derive from SWS. The purpose of the present study was to measure the intelligibility of sinewave replicas of speech signals at the phonetic level, using vowel intelligibility as a starting point.

In a study aimed at testing vocal tract normalization (Ladefoged and Broadbent, 1957) for sinewave utterances, Remez *et al*. (1987) measured the intelligibility of the vowels in *bit, bet, bat*, and *but* when preceded by the phrase *Please say what this word is*. In a control condition—the only condition relevant to the present study—the formant values for the carrier phrase were measured from the same talker who produced the four test words, while in the conditions designed to test vocal tract normalization the formant values were frequency shifted in various ways to simulate different talkers. Vowel intelligibility for the control condition averaged about 60%. This figure is clearly well above the 25% that would be expected by chance, clearly showing that listeners derive phonetic information from SWS without the benefit of higher level sources of information. However, this figure is quite low in relation to the intelligibility of either naturally spoken vowels or vowels generated by a formant synthesizer or the Pattern Playback. For example, Peterson and Barney (1952) reported 94.4% intelligibility for 10 vowel types in /hVd/ syllables naturally spoken by 76 talkers (33 men, 28 women, and 15 children), and Hillenbrand *et al*. (1995) reported 95.4% intelligibility for 12 vowel types in /hVd/ syllables spoken by 139 talkers (45 men, 48 women, and 46 children). Formant-synthesized vowels are also highly intelligible, though less so than naturally spoken vowels. Hillenbrand and Nearey (1999) reported 88.5% intelligibility for formant-synthesized versions of vowels excised from 300 /hVd/ syllables drawn from the Hillenbrand *et al*. recordings. Finally, using the Pattern Playback Ladefoged and Broadbent reported 76% intelligibility for four test words differing in vowel identity (/bɪt/, /bɛt/, /bæt/, /bʌt/).

A straightforward comparison between the sinewave vowel data from Remez *et al*. (1987) and the natural and formant-synthesized speech studies cited above is not possible for obvious reasons. The Remez *et al*. study was not designed as a general test of sinewave vowel intelligibility but rather as an examination of vocal tract normalization. As such, signals were modeled after the recordings of a single talker and only four vowel types were used.

The present study was designed to measure the intelligibility of sinewave vowels more comprehensively using 12 vowel types and stimuli modeled on recordings from a large, diverse group of talkers. A second purpose was to explore the effects of training on the intelligibility of sinewave vowels. As the original Remez *et al*. (1981) study showed, listeners derive much more linguistic information from sinewave replicas simply by being asked to hear the test signals as speech. Both the Remez *et al*. (1987) findings and our own pilot data suggested that the intelligibility of sinewave vowels would be substantially lower than that of the natural utterances used to create them. We therefore tested the effects of three different training procedures that were intended to allow listeners to more clearly apprehend the connection between the odd sounding sinewave replicas and the natural utterances on which they were based. Specifically, all listeners were given an initial test of sinewave vowel intelligibility using 300 stimuli drawn from a large, multitalker vowel database. Listeners were then randomly assigned to one of four conditions: (1) a *feedback* task very similar to the initial vowel intelligibility test, except that listeners were given feedback indicating the vowel category intended by the talker; (2) a *sentence transcription* task in which subjects attempted to transcribe short, simple, and grammatically well-formed sinewave sentences; (3) a task we called *triad* in which subjects listened to a sinewave vowel, followed by the naturally spoken version of that same vowel, followed again by the sinewave vowel; and (4) an irrelevant *control* task in which listeners were asked to judge whether utterances drawn from the /hVd/ database were spoken by men or women.

II. EXPERIMENT 1

A. Methods

1. Stimuli

The test signals used to measure sinewave vowel intelligibility were modeled after 300 utterances drawn from the 1,668 /hVd/ syllables (/i,ɪ,e,ɛ,æ,ɑ,ɔ,o,ʊ,u,ʌ,ɚ/) recorded from 45 men, 48 women, and 46 10- to 12-year-old children by Hillenbrand *et al*. (1995; hereafter H95). The 300 signals were selected at random from the full database, but with the following restrictions: (a) signals showing formant mergers in F_1–F_3 were omitted, (b) signals with identification error rates of 15% or greater (as measured in H95) were omitted, and (c) all 12 vowels were equally represented. The 300 stimuli included tokens from 123 of the original 139 talkers, with 30% of the tokens from men, 36% from women, and 34% from children. This signal set will be referred to as V300. A second set of 180 signals was selected from the H95 database. These signals, which were used in the *feedback* and *triad* training procedures, were selected using the scheme described above, except that signals in the 300-stimulus set were excluded. This signal set will be referred to as V180.

Sinewave replicas of the signals in V300 and V180 were generated from the hand edited formant tracks measured in H95. The method involved extracting peaks from LPC spectra every 8 ms, followed by hand editing *during the vowel only* using a custom interactive editing tool.[1] Each sinewave replica was generated as the sum of three sinusoids that followed the measured frequencies and amplitudes of F_1–F_3 during the vowel portion of the /hVd/ syllable. The signals were synthesized at the same 16 kHz sample rate that had been used for the original digital recordings. Following synthesis all signals were scaled to a common rms amplitude.

Sinewave sentences for the *sentence transcription* task were synthesized using 50 sentences drawn at random from the 250-sentence Hearing in Noise Test (HINT) recordings (Macleod and Summerfield, 1987; Nilsson *et al.*, 1994). The utterances in this database are brief, syntactically and semantically simple sentences (e.g., "Her shoes were very dirty.") that are carefully spoken by a single adult male talker. Because generating sinewave replicas of these sentences from hand-edited formant tracks would have been quite time consuming, a fully automated method was used to generate the test signals from unedited spectral envelope peaks. The method is a broadband version of the narrowband sinusoidal synthesis method described in Hillenbrand *et al.* (2000), where it is described in greater detail. Briefly, as illustrated in Fig. 2, the major signal processing steps for each 10-ms frame include: (1) a 32-ms Hamming-windowed Fourier spectrum; (2) calculation of a masking threshold as the 328 Hz Gaussian-weighted running average of spectral amplitudes in the Fourier spectrum; (3) subtraction of the masking threshold from the Fourier spectrum, with values below the masking threshold set to zero, a process which has the effect of emphasizing high energy regions of the spectrum (especially formants) at the expense of valleys and minor peaks (see Hillenbrand and Houde 2003, for a discussion); (4) calculation of the smoothed envelope as the 200 Hz Gaussian-weighted running average of the masked spectrum; and (5) extraction of spectral peak frequencies and amplitudes from the envelope. Peaks are extracted for both voiced and unvoiced frames, and there is no limit on the number of peaks per frame, which average about five. Synthesizing a sinewave replica is simply a matter of calculating then summing sinusoids at the measured frequencies and amplitudes of each peak, with durations equal to the frame rate (10 ms in the present case). To avoid phase discontinuities at the boundaries between frames, spectral peaks must be tracked from one frame to the next. If the tracking algorithm determines that a given spectral peak is continuous from one frame to the next, the frequency and amplitude of the peak are linearly interpolated through the frame, and the starting phase of the sinusoid in frame $n+1$ is adjusted to be continuous with the ending phase of the sinusoid in frame n. On the other hand, if the algorithm determines that a spectral peak in frame n does not continue into frame $n+1$, the amplitude of the sinusoid is ramped down to zero. Similarly, if a spectral peak is found in a given analysis frame that is not continuous with a peak in the previous frame, the amplitude of the sinusoid is ramped up from zero to its measured amplitude in the current frame. The HINT sentences were synthesized at 16 kHz and scaled to a common rms value. Our informal impression was that sentences that were synthesized from unedited spectral peaks using this method were as intelligible as those generated from edited formants. In experiment 2 we will report results showing that these SW sentences are, in fact, quite intelligible.

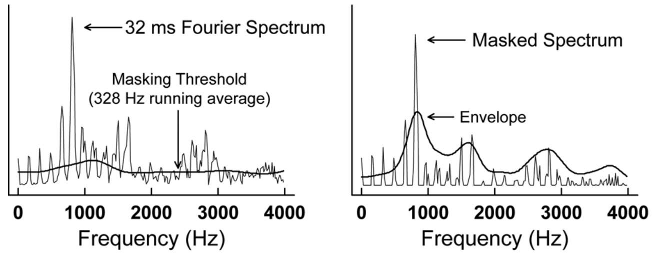

FIG. 2. Signal processing steps used in the automated sinewave synthesis technique (see text).

2. Subjects and procedures

Seventy-one listeners were recruited from students enrolled in an introductory phonetics course. The students passed a pure-tone hearing screening (25 dB at octave frequencies from 0.5–4 kHz) and were given bonus points for their participation. The listeners were drawn from the same geographic regions as the talkers. The great majority of the listeners were from southern Michigan, with others primarily from neighboring areas such as northern Indiana, northwest Ohio, and northeast Illinois. All subjects participated in the main 300-stimulus sinewave vowel intelligibility test. Using general-purpose experiment-control software (Hillenbrand and Gayvert, 2003), sinewave vowels from the V300 set were presented to listeners for identification in random order, scrambled separately for each listener. The stimuli were low-pass filtered at 7.2 kHz, amplified, and delivered free field in a quiet room over a single loudspeaker (Paradigm Titan v.3) positioned about one meter from the listener's head at a level averaging about 75 dBA. Subjects used a mouse to select one of 12 buttons labeled with a phonetic symbol and a key word (*heed, hid, head*, etc.). Listeners were allowed to replay the stimulus before making a response. Feedback was not provided. The sinewave vowel intelligibility test was preceded by a brief, 24-trial practice session using naturally spoken versions of the 12 vowels.

Following the sinewave vowel intelligibility test, listeners were randomly assigned to one of four conditions: (a) *feedback* (N = 19), (b) *sentence transcription* (N = 18), (c) *triad* (N = 16), or (d) an irrelevant control task involving the judgment of speaker sex from /hVd/ syllables (N = 18). Procedures for the *feedback* condition were identical to the vowel intelligibility test, with two exceptions: (1) following the listener's response, one of the 12 buttons was blinked briefly to indicate the correct response, and (2) stimuli from the V180 stimulus set were used. In the *sentence transcription* task, sinewave replicas of 50 randomly ordered HINT sentences were presented to listeners, who were asked to transcribe each utterance by typing in a text entry box on the screen. The *triad* training procedure was identical to the

TABLE I. Confusion matrix for sinewave vowels in experiment 1, prior to training. The response means given in the last row indicate the percentage of trials in which a given vowel was used as a listener response. Each vowel was presented on 8.3% of the trials.

		Vowel identified by listener											
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɚ/
Vowel intended by talker	/i/	**76.7**	3.9	3.1	1.4	0.5	0.5	0.2	1.3	2.3	9.3	0.9	0.0
	/ɪ/	12.8	**44.7**	2.3	8.5	2.9	1.6	0.7	1.7	4.4	12.8	5.6	2.0
	/e/	40.8	2.8	**41.0**	1.8	0.8	0.8	0.4	2.1	1.5	6.4	0.6	1.0
	/ɛ/	2.6	17.1	8.2	**30.4**	9.1	3.4	1.7	7.3	6.3	6.0	5.6	2.2
	/æ/	1.5	8.3	6.9	16.3	**36.9**	2.4	3.8	4.7	2.3	5.3	2.1	9.6
	/ɑ/	0.6	0.8	2.9	2.5	5.5	**33.1**	15.5	5.0	5.0	2.1	17.7	9.3
	/ɔ/	0.1	0.3	0.4	0.5	1.1	12.8	**52.3**	19.4	3.5	2.0	6.3	1.2
	/o/	0.0	0.0	0.1	0.2	0.1	0.6	0.6	**69.7**	2.6	24.7	0.7	0.7
	/ʊ/	0.2	1.2	1.4	1.3	0.7	1.2	1.0	1.5	**66.7**	11.3	10.9	2.5
	/u/	0.1	0.1	0.1	0.1	0.0	0.2	0.1	12.9	8.1	**75.9**	2.1	0.5
	/ʌ/	0.2	1.0	1.1	2.1	1.4	3.7	3.5	7.2	33.5	5.4	**39.8**	1.3
	/ɚ/	0.3	0.1	0.7	0.5	0.2	0.2	0.1	0.7	1.4	2.7	0.8	**92.4**
Response means:		11.3	6.9	5.7	5.5	4.9	5.0	6.7	11.1	11.5	13.7	7.8	10.2

sinewave vowel intelligibility test, except that: (1) the stimuli for each trial consisted of a sinewave vowel, the naturally spoken version of that vowel, followed again by the sinewave vowel, (2) the listener's response was followed by feedback in the form of a button blink, and (3) signals from the V180 stimulus set were used. Listeners were free to respond any time after the start of the first stimulus, although listeners typically waited to hear all of the signals. (Listeners' performance on the triad task is not really relevant because subjects heard both the SW and the highly intelligible natural version of the each stimulus on each trial.) The training sessions were self-paced and took, on average, about 30 min to complete.

B. Results and discussion

Vowel intelligibility, averaged across all 71 listeners prior to training, was 55.0%, a figure that is some 40 percentage points lower than that for naturally spoken versions of the same vowels (Hillenbrand and Nearey, 1999). Although the sinewave vowels are much less intelligible than naturally spoken vowels, the 55.0% intelligibility figure is substantially greater than the 8.3% that would be expected by chance, clearly showing that a good deal of phonetic information is conveyed by the sinewave replicas. Variability across listeners was quite large, with a standard deviation of 12.7, a coefficient of variation of 0.23, and a range of 21.3%–80.7%. Large intersubject variability has characterized sinewave speech findings from the start (e.g., Remez *et al.*, 1981; Remez *et al.*, 1987).

The confusion matrix for the sinewave vowels, displayed in Table I, shows some features in common with comparable data for naturally spoken vowels, especially the high identification rates for /i/, /u/, and /ɚ/ relative to the other vowels, a feature that is virtually always seen in labeling data for naturally spoken vowels (e.g., Peterson and Barney, 1952; Hillenbrand *et al.*, 1995; Hillenbrand and Nearey, 1999). There are, however, many differences. Of special note are the very low identification rates for /ɛ/, /ɑ/, /ʌ/, and /æ/, and the many confusions among phonetically dissimilar vowels that are almost never observed for naturally spoken vowels presented in quiet. A correlation was calculated relating the arcsine-transformed percent correct values for the 12 individual vowel categories for the sinewave stimuli from the present study with comparable figures for the naturally spoken versions of the same vowels from Hillenbrand and Nearey (1999). The correlation was significant but not strong ($r = 0.44$, $p < 0.05$).

Figure 3 compares vowel intelligibility scores prior to and following training under the four conditions. It can be seen that post-training performance is uniformly higher than pre-training, with the largest increment for the *triad* condition (10.9 percentage points) and the smallest increment for the control condition (2.5 percentage points). A two-way ANOVA on arcsine-transformed intelligibility scores showed a significant effect for Training (i.e., pre- versus post-training, [F(1,67) = 96.1, $p < 0.0001$], no effect for Condition (i.e., control versus feedback, etc.), and a

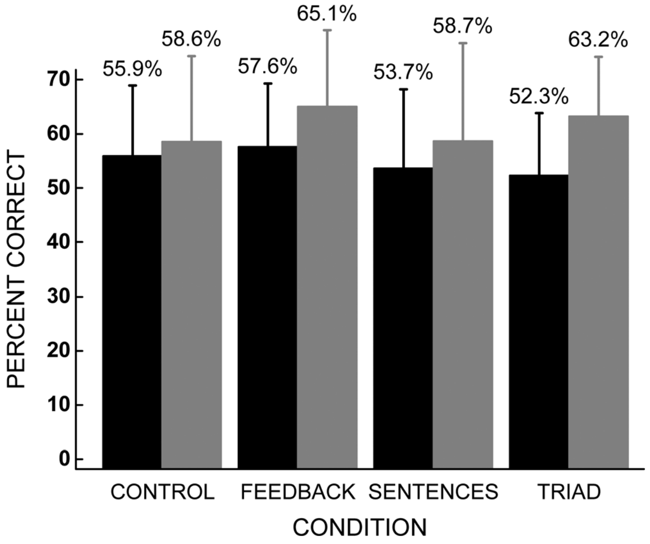

FIG. 3. Sinewave vowel intelligibility prior to (dark bars) and following (shaded bars) a control condition and three training conditions. Error bars indicate one standard deviation.

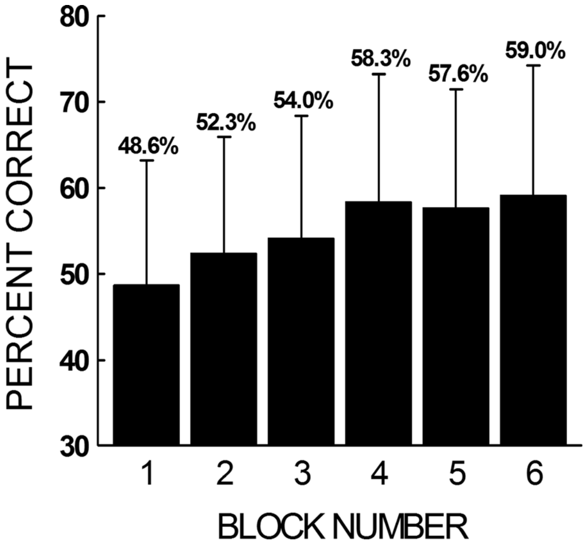

FIG. 4. Sinewave vowel intelligibility as a function of block number for the 300-trial pre-test. Block 1 consists of trials 1–50, block 2 consists of trials 51–100, etc. Error bars indicate one standard deviation.

significant interaction [F(3,67) = 6.3, $p < 0.001$]. Planned comparisons using the Holm (1979) step-down procedure showed significant differences between pre- and post-training vowel intelligibility scores for all four conditions, including the irrelevant control condition. The source of the interaction is readily evident in Fig. 3, which shows that the training effect is larger for some conditions than others. Planned comparisons on difference scores (post-training minus pre-training) using the Holm procedure showed significantly larger training effects for *triad* versus *control*, *feedback* versus *control*, and *triad* versus *sentence transcription*. The *sentence transcription* versus *control* comparison fell just short of significance.

The intelligibility scores for all four post-training conditions were significantly higher than the pre-test scores. The small but reliable effect for the control condition (an irrelevant task unrelated to either SWS or vowel identification) shows that listeners' performance improves as a result of simple exposure to the SW vowels in the pre-test and/or repeated attempts to identify the signals. This is consistent with analyses of pre-test results showing that performance steadily improves as the experiment progresses. Figure 4 shows pre-test performance as a function of block number, with the 300 trials arbitrarily divided into blocks of 50 trials. The improvement in performance with increasing block number is highly significant [F(5,70) = 20.0, $p < 0.00001$], although it appears to plateau at block 4. *Post hoc* tests (Holm, 1979) comparing adjacent blocks showed statistically reliable differences only between blocks 1 and 2 and blocks 3 and 4.

The magnitude of the training effect differed considerably across the three non-control training conditions. In absolute terms, the two most effective training conditions—*feedback* and *triad*—involved specific exposure to sinewave vowels and called for judgments of the identity of those vowels. Further, the most effective procedure exposed subjects to both sinewave and natural versions of the same utterance, establishing an immediate and direct connection between the two types of signals. It is important to note that the relatively weak training effect for the *sentence transcription* task cannot be attributed to the poor intelligibility of the SW sentences. Word-level transcription accuracy averaged across the 18 listeners in this condition was 89.6% (SD = 8.4).

These details aside, the single most striking aspect of the training results concerns the modest effectiveness of all of the training procedures in absolute terms. Listeners in *triad*, the most effective of the training conditions, identified 300 sinewave vowels in the pre-test, with specific instructions to hear these odd-sounding signals as vowels, followed by 180 trials in which the sinewave and natural versions of each signal were presented back-to-back. Yet following this experience, sinewave vowel intelligibility remained low—by some 32 percentage points—relative to naturally spoken versions of the same vowels. Prior to training, then, listeners had a good deal of difficulty identifying sinewave vowels, and although performance was reliably better following training, listeners still had a good deal of difficulty identifying these vowels.

III. EXPERIMENT 2

The purpose of the training procedures that were tested in experiment 1 was to determine whether the intelligibility of SW vowels would improve if listeners were provided with some minimal training using SW speech. While all of the training methods produced statistically reliable improvements in intelligibility, the effects were modest in absolute terms. The purpose of experiment 2 was to measure the effects of more extensive training. A separate group of listeners was tested prior to and then following five 180-trial blocks of training using a slight variation of the *triad* method, the most effective of the training techniques tested in experiment 1.

A. Methods

Stimuli and procedures for the pre- and post-test were identical to those described for experiment 1. The pre-test was followed by five 15–20 min training sessions using the V180 stimulus set and a modified *triad* procedure. In this variation subjects listened to and identified the SW version of each of the V180 signals, followed by feedback in the form of a button blink. If the vowel was identified correctly the program simply cycled to the next trial. However, if the vowel was misidentified, listeners would hear sinewave, natural, and sinewave versions of the stimulus. As with experiment 1, the training trials were followed by a post-test using the V300 stimulus set.

Twelve listeners were recruited from an undergraduate hearing science class. The listeners had taken an introductory phonetics course the previous semester. As with experiment 1, (1) the listeners had normal hearing (25 dB or better, 0.5–4 kHz), (2) they were drawn from the same geographic region as the talkers, and (3) they were given bonus points

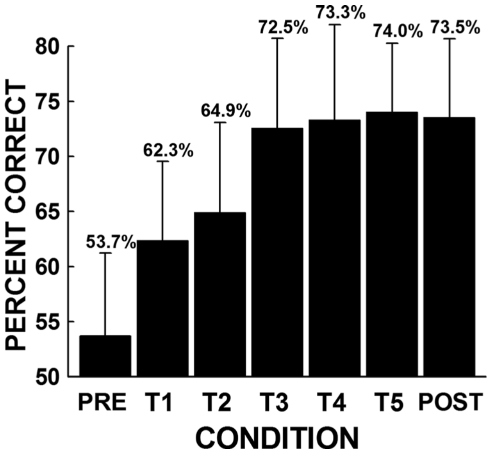

FIG. 5. Sinewave vowel intelligibility for a pre-test, five blocks of training using a *triad* procedure, and a post-test. Error bars indicate one standard deviation.

for their participation. None of these listeners had participated in experiment 1.

B. Results and discussion

Results for the pre-test, the five training conditions (T1–T5), and the post-test are shown in Fig. 5. It can be seen that there is an immediate improvement in performance of 8.6 percentage points from the pre-test to T1. Performance continues to improve through T3, with little improvement following that. These visual observations were confirmed by a repeated-measures ANOVA, which showed a highly significant effect for Condition [F(6,77) = 12.0, $p < 0.0001$]. Planned comparisons using the Holm (1979) procedure showed reliable differences between the pre-test and T1 and between T2 and T3. No significant improvements were seen for the other pairs of adjacent conditions. Improvement on the V300 stimuli from pre-test to post-test was a very substantial 19.8 percentage points.

In summary, findings from experiment 2 show that additional training beyond the single block of trials offered in experiment 1 produces substantial improvements in vowel intelligibility; however, performance asymptotes at about 73%–74% after three training sessions. Even after this more extensive training, the identification rate remains well below the 95.4% intelligibility reported for naturally spoken versions of the same signals (Hillenbrand and Nearey, 1999).

IV. PATTERN RECOGNITION

In their original article on sinewave speech, Remez *et al*. (1981) noted that SWS is, "… a deliberately abstract representation of the time-varying spectral changes of the naturally produced utterance, although in local detail it is unlike natural speech signals …" (p. 948). They went on to observe that SWS, "… consists of none of those distinctive acoustic attributes that are traditionally assumed to underlie speech perception. … For example, there are no formant frequency transitions, which cue manner and place of articulation; there are no steady-state formants, which cue vowel color and consonant voicing; and no fundamental frequency changes, which cue voicing and stress. … [The] short-time spectral cues, which depend on precise amplitude and frequency characteristics across the harmonic spectrum, are absent from these tonal stimuli. … [If] listeners are able to perceive the tones as speech, then we may conclude that traditional speech cues are themselves approximations of second-order signal properties to which listeners attend" (p. 948). The ability shown by listeners in the present study to identify sinewave vowels (though imperfectly), in spite of the many differences in spectral detail between the sinewave replicas and the natural utterances upon with they were modeled, might seem to rule out template matching as the underlying pattern recognition principle and argue instead for mental computations involving more abstract representations, perhaps corresponding to the "second-order signal properties" mentioned by Remez *et al*. The purpose of the work presented in this section is to test this idea by attempting to recognize SW vowels using a template-based pattern recognizer whose templates are derived empirically from naturally spoken vowels.

The template-matching algorithm that was used for this work was the narrow-band pattern-matching model of vowel perception described in Hillenbrand and Houde (2003). A full description can be found in that paper, but briefly, each of 12 vowel categories (/i,ɪ,e,ε,æ,ɑ,ɔ,o,ʊ,u,ʌ,ɚ/) is represented by a sequence of smooth spectral shape templates that are empirically derived by averaging narrow band Fourier spectra from naturally spoken vowels of a given type sampled at comparable time points throughout the course of the vowel. The standard version of the model samples the vowels at five equally spaced points between 15% and 75% of vowel duration (Fig. 6). Separate template sequences are used to define vowel templates for men, women, and children.

The signal processing for template construction begins with the calculation of a 64 ms, Hamming windowed Fourier spectrum, using linear frequency and amplitude scales (Fig. 7). The next step is spectrum level normalization, designed

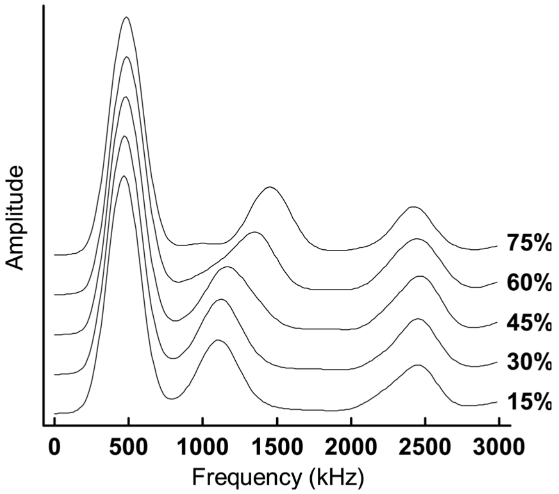

FIG. 6. Sequence of five spectral shape templates defining the [ʊ] vowel category for adult male talkers (adapted from Hillenbrand and Houde, 2003). Templates were created at five time slices equally spaced between 15% and 75% of vowel duration.

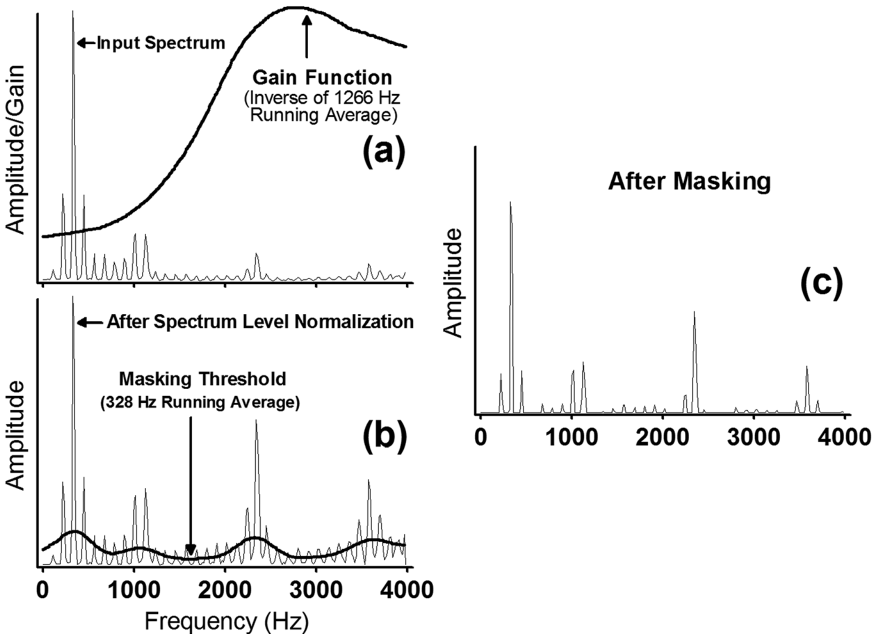

FIG. 7. Signal processing steps used to create vowel templates (adapted from Hillenbrand and Houde, 2003).

to flatten the spectrum, reducing as much as possible within-vowel-category differences in formant amplitude relations, which contribute very little to the perception of vowel quality (Klatt, 1982). This is done by multiplying the Fourier spectrum by a gain function consisting of the inverse of the Gaussian-weighted running average of spectral amplitudes, using a 1266 Hz averaging window (Fig. 7, panels a and b). The next step was designed to enhance spectral peaks—both formants and harmonics—at the expense of minor spectral details in between peaks. This was done by calculating a threshold function as the 328 Hz Gaussian-weighted running average of amplitudes in the level-normalized spectrum. The threshold function is simply subtracted from the level-normalized spectrum, with values below the threshold being set to zero (Fig. 7, panels b and c). In previous work we have referred to this step variously as a *thresholding* operation and as *masking* operation (see Hillenbrand and Houde, 2003, footnote 2). Separate template sequences for men, women, and children are derived from these level-normalized, masked narrow band spectra simply by averaging like vowels at comparable time points (i.e., 15% of vowel duration, 30% of vowel duration, etc.), followed by light smoothing using a 172 Hz Gaussian-weighted running average. For the present study, each template was constructed by averaging approximately 40 tokens of each vowel at each time slice using signals from the H95 database. Omitted from the database for the purposes of template creation were 192 utterances with identification error rates of 15% or greater.

Prior to calculating the token-template distance, the input spectrum is processed by the level normalization and masking operations described above. The token-template distance is simply the sum of the channel-by-channel absolute differences between the token spectrum and the smooth template, divided by the sum of the amplitudes in the template (i.e., an amplitude-normalized city-block distance). It is important to note that these differences are calculated for all 256 channels, despite the fact that the differences are relevant only at the harmonic frequencies that define the envelope shape. Differences at frequencies that are remote from harmonics are large and irrelevant. An important assumption of the model is that these irrelevant inter-harmonic differences will be about equally large to all vowel templates, resulting in a roughly constant source of noise (cf. de Cheveigné and Kawahara, 1999, for a related algorithm which compares the token and template only at harmonic frequencies). It might be noted that this assumption, which has proven to be reasonable in earlier work, is put to an extreme test in the case of sinusoidal vowels in which the envelope shape is defined by just three of the 256 frequency components.

At each of the five time slices, the spectral-distance algorithm produces a 12-element vector holding city-block distances between the input spectrum and templates for each of the twelve vowels categories (Fig. 8). The final distance vector that is used for recognition is the weighted average of the five individual vectors. A weight of 0.6 was used for the distance calculated from the last slice, which shows the influence of the final consonant, and weights of 1.0 for distances computed from the remaining slices. The vowel is recognized as the category producing the smallest token-template distance in the average distance vector.

A. Results and discussion

The confusion matrix for the narrow band model, averaged across the three talker groups, is shown in Table II. The average correct classification rate was 78.3%, with rates of 82.2%, 76.0%, and 76.8% for men, women, and children, respectively. The 78.3% overall classification rate for the model is much higher than the 55.0% and 53.7% pre-training intelligibility rates shown by listeners in experiments 1 and 2, respectively, but only about 5 percentage points higher than the 73.5% recognition rate shown by listeners following the more extensive training used in experiment 2. As with human listeners, the model is much more accurate in

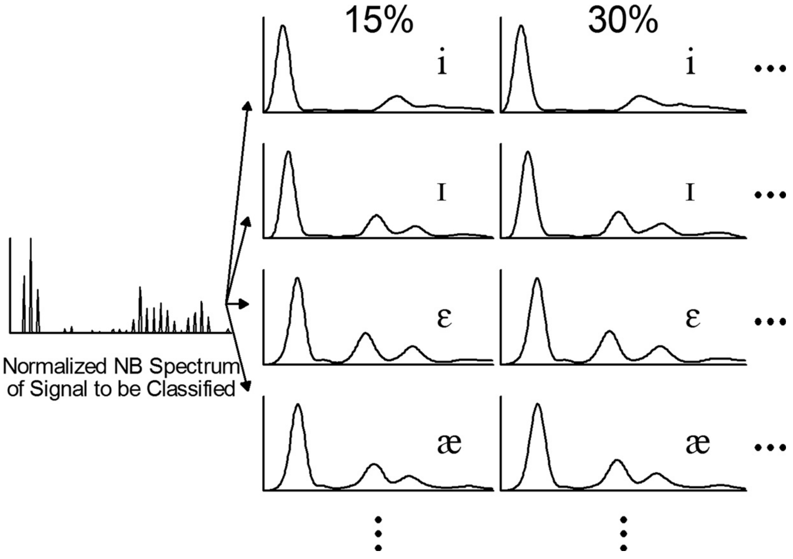

FIG. 8. Sketch of the recognition method used by the narrow band model. The normalized narrow band spectrum computed at 15% of vowel duration is compared to the 12 vowel templates computed at the same time point (only four of which are shown here). The narrow band input spectrum at 30% of vowel duration (not shown) is then compared to the 12 vowel templates computed at 30% of vowel duration, and so on (from Hillenbrand and Houde, 2003).

classifying naturally spoken than SW vowels. The model correctly classified 91.4% of naturally spoken vowels in the full, 1668-utterance H95 database (men: 92.0%, women: 92.1%, children: 90.0%), and 93.7% of the 300-vowel subset that was used in the present study (men: 95.6%, women: 95.4%, children: 90.2%).[2] For comparison with the model, Table III shows the confusion matrix for listeners in the post-training condition of experiment 2. The confusion matrices in Table II and III are by no means identical, but there are many points of similarity. With some notable exceptions (especially model confusions between /i/ and /u/) the model tends to make the same kinds of errors as the listeners. The cell-by-cell correlation between the two confusion matrices is 0.96.

The main conclusion to be drawn from these findings is that, in spite of the many differences in spectral detail between sinewave and natural speech, the sinewave speech phenomenon does not rule out relatively straightforward pattern recognition schemes based on template matching. Of course these findings do not compel an interpretation based on template matching. It bears repeating that the reasonable classification performance of the model (24 percentage points higher than our listeners prior to training, and more-or-less on par with the listeners following the more extensive training used in experiment 2) is based on templates created from natural rather than sinewave speech, and that only three of the values in each 256-point city-block distance vector were relevant to measuring the similarity between the sinewave input spectrum and the template. We attribute the reasonable performance of the model mainly to the steps that were taken (spectrum level normalization and masking) to emphasize differences in formant frequencies at the expense of other aspects of spectral shape. It is, of course, the formant pattern and only the formant pattern which ties the original signal to the sinewave replica.

V. SUMMARY AND GENERAL DISCUSSION

The primary purpose of this study was to measure the intelligibility of sinewave replicas of speech signals explicitly at the phonetic level. Vowel intelligibility was chosen as an arbitrary starting point. The test signals were sinusoidal

TABLE II. Confusion matrix for the narrow band vowel recognition algorithm.

		Vowel classified by the narrow band model											
		/i/	/ɪ/	/e/	/ε/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɚ/
Vowel intended by talker	/i/	**80.2**	1.6	10.3	0.0	0.0	0.0	0.0	2.4	0.0	5.6	0.0	0.0
	/ɪ/	2.2	**89.2**	6.5	0.0	0.0	0.0	0.0	0.7	1.4	0.0	0.0	0.0
	/e/	8.1	6.5	**82.3**	0.0	0.0	0.0	0.0	2.4	0.0	0.8	0.0	0.0
	/ε/	0.0	1.4	0.0	**87.8**	7.2	0.0	2.2	0.0	0.0	0.0	0.7	0.7
	/æ/	0.0	1.5	0.0	25.0	**71.3**	0.0	2.2	0.0	0.0	0.0	0.0	0.0
	/ɑ/	0.0	0.0	0.0	0.7	1.5	**65.4**	15.4	0.0	0.0	0.0	16.9	0.0
	/ɔ/	0.0	0.0	0.0	2.3	1.5	3.8	**79.7**	0.0	0.0	0.0	12.8	0.0
	/o/	1.5	0.0	10.2	0.7	0.7	0.0	1.5	**78.1**	0.7	5.8	0.7	0.0
	/ʊ/	0.7	2.9	2.2	0.0	0.0	0.0	0.0	0.7	**84.9**	5.8	2.9	0.0
	/u/	29.2	0.7	10.9	0.0	0.0	0.0	0.0	4.4	3.6	**51.1**	0.0	0.0
	/ʌ/	0.0	0.0	0.0	1.5	0.7	0.0	3.7	0.7	14.7	0.0	**78.7**	0.0
	/ɚ/	0.0	1.7	5.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**93.2**
Response means:		10.2	8.8	10.6	9.8	6.9	5.8	8.7	7.5	8.8	5.8	9.4	7.8

TABLE III. Confusion matrix for the post-training condition of experiment 2.

		Vowel identified by the listener											
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɚ/
Vowel intended by talker	/i/	**77.7**	3.7	9.7	0.3	0.3	0.0	0.0	0.7	1.0	6.7	0.0	0.0
	/ɪ/	4.0	**73.9**	6.6	7.9	1.0	0.0	0.0	0.0	2.0	4.0	0.3	0.3
	/e/	9.9	2.6	**85.8**	0.3	0.0	0.0	0.3	0.0	0.0	0.1	0.0	0.0
	/ɛ/	0.0	10.0	4.0	**67.1**	5.6	2.0	1.7	2.3	1.7	1.3	3.3	1.0
	/æ/	0.0	3.3	1.0	15.7	**70.0**	2.7	1.7	1.7	2.0	0.7	0.3	1.0
	/ɑ/	0.3	0.3	0.0	5.3	6.0	**52.8**	7.6	3.7	2.0	0.0	18.9	3.0
	/ɔ/	0.0	0.0	0.0	0.0	0.0	20.0	**64.0**	8.7	1.0	1.0	5.0	0.0
	/o/	0.0	0.0	0.0	0.0	0.0	0.0	0.3	**90.7**	0.7	7.3	0.7	0.3
	/ʊ/	0.0	0.7	0.7	0.3	0.0	0.0	0.0	0.3	**78.7**	6.7	9.0	1.0
	/u/	0.0	0.3	0.3	0.0	0.0	0.0	0.0	19.0	5.0	**75.0**	0.0	0.3
	/ʌ/	0.0	1.0	0.0	4.6	1.0	2.3	0.3	8.6	24.1	3.3	**54.8**	0.0
	/ɚ/	0.0	0.0	4.0	0.7	0.0	0.0	0.0	0.0	1.7	1.0	0.7	**92.0**
Response means:		7.7	8.9	9.3	8.5	7.0	6.7	6.3	11.3	10.0	8.9	7.8	8.2

replicas of vowels excised from /hVd/ syllables. Vowels were used so that listeners would be unable to make use of higher level knowledge sources such as syntax, semantics, and the lexicon. The intelligibility of the SW vowels was 55.0%, with a great deal of variability across listeners. This figure is many times greater than the 8.3% that would be expected based on chance, showing that there is a good deal of phonetic information in the sinusoidal replicas. On the other hand, the figure is roughly 40 percentage points lower than the intelligibility of naturally spoken versions of the same test signals. Even the 55.0% figure for the pre-test is an overestimate of listeners' ability to identify SW vowels at initial exposure. A trend analysis showed performance at 48.6% for the first block of 50 trials, with performance rising fairly steadily to 59.0% for the final block of trials (Fig. 4).

The modest intelligibility of sinewave vowels cannot be attributed to the loss of potentially important coarticulation cues resulting from the use of vowels that were excised from /hVd/ syllables. Hillenbrand and Nearey (1999) compared the intelligibility of both natural and formant synthesized /hVd/ syllables with that of vowels excised from those syllables. Vowels excised from natural /hVd/ syllables were slightly less intelligible (by 2.3 percentage points) than the full syllables, and no difference was observed between excised and non-excised formant-synthesized utterances. Another finding from Hillenbrand and Nearey that is relevant to the present study concerns intelligibility tests using natural and formant-synthesized versions of the V300 signals. The naturally spoken signals were significantly more intelligible than the formant-synthesized signals (95.4% versus 88.5%), showing that phonetically relevant information is lost when vowels are reduced to a formant representation. However, formant-synthesized versions of the V300 signals from Hillenbrand and Nearey were 33.5 percentage points more intelligible than SW versions of the same signals that were tested in the present study, this despite the fact that both types of signals were driven by the same underlying formant information.

Given the strange, unfamiliar sound quality of sinusoidal speech, we considered the possibility that a modest amount of training might produce substantial improvements in performance. Three training procedures were tested: *feedback, sentence transcription*, and a *triad* method in which subjects listened to a sinewave vowel, followed by the naturally spoken version of that vowel, followed again by the sinewave version. All three training methods produced statistically reliable improvements in SW vowel intelligibility. In fact, a small but reliable improvement was seen following a control task that did not involve either sinusoidal speech or vowel intelligibility. This shows that simply attempting to identify the SW vowels in the pre-test, even in the absence of feedback, is enough to produce an improvement in performance. The *sentence transcription* task produced a reliable but rather small improvement in performance of just 5.0 percentage points—only a few points better than the control task and not significantly different from it. This training effect was relatively small despite the fact that listeners had little difficulty transcribing the sinewave sentences, with transcription accuracy averaging slightly under 90%. Listeners in this condition, then, were clearly able to hear sinewave replicas as speech, yet this experience in transcribing SW sentences transferred only weakly to the task of identifying SW vowels. It may be that hearing SW sentences as speech does not guarantee that listeners will find it easy to hear relatively brief, isolated SW vowels as speech. The two most effective methods, *feedback* and *triad*, focused specifically on SW vowels, and *triad*, the most effective method, allowed listeners to hear SW and naturally spoken versions of the same signal in close succession.

All of the training methods used in experiment 1 produced statistically reliable effects, but what we found most striking about these results is that all of the training effects were of modest magnitude. Following training, performance remained about 30–37 percentage points below that for naturally spoken versions of the same vowels (Hillenbrand and Nearey, 1999). The training results in experiment 1, however, were based on a single session lasting about 30 min. Experiment 2, which used a variation of the *triad* method, was designed to determine whether listeners would benefit from more extensive training. The results were quite clear.

Performance improved over the first three of five 180-trial training blocks. There was no further benefit of additional training, with performance leveling off at about 73%–74%, approximately 22 percentage points below that for naturally spoken versions of the same utterances.

One other aspect of the findings that merits brief mention concerns the fact that highly intelligible SW sentences were generated using a method that did not involve explicit formant tracking. To our knowledge, all previous work on the perception of sinusoidal sentences has used utterances generated from formant tracks. The sinewave versions of the HINT sentences used in the present study, on the other hand, were generated using a fully automated method that was driven by unedited envelope peaks. The reasonable intelligibility of the sinusoidal HINT sentences used here is consistent with work using a damped sinewave synthesizer (a method similar in broad principle to formant synthesis which produces speech that is far more natural sounding than SWS) showing that highly intelligible speech can be generated from unedited envelope peaks (Hillenbrand *et al.*, 2006).

A template-based vowel recognition algorithm, which was trained on naturally spoken vowels, classified the same SW vowels that were used in the two perception experiments with 78.3% accuracy. This figure is much higher than the 55% recognition rate shown by human listeners prior to training and fairly similar to the 73%–74% recognition rate shown by listeners following the more extensive training used in experiment 2. These modeling results show that, in spite of the many differences in acoustic detail between natural and sinewave speech, the ability of listeners to recognize SWS does not rule out an underlying recognition process that is based on template matching.

Listeners had a much easier time recognizing sinewave replicas of meaningful and grammatically well-formed sentences than they did recognizing isolated vowels. A perfectly obvious explanation for this is that isolated vowels limit listeners to recognition processes at the acoustic–phonetic level while sentences allow them to make use of rich sources of knowledge at higher levels of the language processing system. However, this explanation does not exhaust the possibilities. In addition to the effects of linguistic knowledge, it is possible that longer sentence-length utterances give listeners an opportunity to make perceptual accommodations to the strange sounding sinewave utterances. There is evidence suggesting that this is, in fact, the case. Hillenbrand *et al.* (2008) asked listeners to identify sinusoidal replicas of /hVd/ syllables either in isolation or preceded by a brief SW carrier phrase ("The next word in the list is …"). Vowel intelligibility averaged 53.5% under the isolated vowel condition but 73.1% when the same utterances were preceded by the SW carrier phrase. Additional findings show that some of the carrier phrase effect is related to listeners accommodating to the characteristics of the individual talker, but there is also clear evidence that more is involved than talker normalization. Exactly what subjects are learning while listening to the sinewave carrier phrase is as yet unclear.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Institutes of Health (R01-DC01661) to Western Michigan University. We are grateful to two anonymous reviewers for very helpful comments on an earlier draft.

[1] Vowels were excised between the onset of voicing and the onset of the /d/ occlusion interval, signaled primarily by a sharp decrease in the amplitudes of formants above F1. Consequently, what is referred to here as the "vowel" included the formant transitions into the final /d/ (see Hillenbrand and Nearey, 1999, Fig. 2).

[2] We assume that classification rates are higher for the V300 signals because signals that were poorly identified by listeners were removed from this subset (see Hillenbrand and Houde, 2003, for some evidence).

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

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am., **24**, 597–606.

Hillenbrand, J. M., Clark, M. J., and Houde, R. A. (**2000**). "Some effects of duration on vowel recognition," J. Acoust. Soc. Am. **108**, 3013–3022.

Hillenbrand, J. M., Clark, M. J., Houde, R. A., Hillenbrand, M. W., and Hillenbrand, K. S. (**2008**). "Perceptual accommodation to sinewave speech," J. Acous. Soc. Am. **124**, 2435.

Hillenbrand, J. M., and Gayvert, R. A. (**2005**). "Open source software for experiment design and control," J. Speech Lang. Hear. Res. **48**, 45–60.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Hillenbrand, J. M., and Houde, R. A. (**2003**). "A narrow band pattern-matching model of vowel perception," J. Acoust. Soc. Am. **113**, 1044–1055.

Hillenbrand, J. M., Houde, R. A., and Gayvert, R. A. (**2006**). "Speech perception based on spectral peaks versus spectral shape," J. Acoust. Soc. Am. **119**, 4041–4054.

Hillenbrand, J. M., and Nearey, T. M. (**1999**). "Identification of resynthesized /hVd/ syllables: Effects of formant contour," J. Acoust. Soc. Am. **105**, 3509–3523.

Holm, S. (**1979**). "A simple sequentially rejective multiple test procedure," Scand. J. Stat. **6**, 65–70.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 13–33.

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

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 88–104.

Macleod, A., and Summerfield, Q. (**1987**). "Quantifying the contribution of vision to speech perception in noise," Br. J. Audiol. **21**, 131–141.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

Peterson, G., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

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

Remez, R. E., Rubin, P. E., Nygaard, L. C., and Howell, W. A. (**1987**). "Perceptual normalization of vowels produced by sinusoidal voices," J. Exp. Psychol.: Hum. Percept. Perform. **13**, 40–61.