The role of contrasting temporal amplitude patterns in the perception of speech

Eric W. Healy[a] and Richard M. Warren
Department of Psychology, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, Wisconsin 53201

(Received 24 September 2001; revised 27 November 2002; accepted 16 December 2002)

Despite a lack of traditional speech features, novel sentences restricted to a narrow spectral slit can retain nearly perfect intelligibility [R. M. Warren *et al.*, Percept. Psychophys. **57**, 175–182 (1995)]. The current study employed 514 listeners to elucidate the cues allowing this high intelligibility, and to examine generally the use of narrow-band temporal speech patterns. When $\frac{1}{3}$-octave sentences were processed to preserve the overall temporal pattern of amplitude fluctuation, but eliminate contrasting amplitude patterns within the band, sentence intelligibility dropped from values near 100% to values near zero (experiment 1). However, when a $\frac{1}{3}$-octave speech band was partitioned to create a contrasting pair of independently amplitude-modulated $\frac{1}{6}$-octave patterns, some intelligibility was restored (experiment 2). An additional experiment (3) showed that temporal patterns can also be integrated across wide frequency separations, or across the two ears. Despite the linguistic content of single temporal patterns, open-set intelligibility does not occur. Instead, a contrast between at least two temporal patterns is required for the comprehension of novel sentences and their component words. These contrasting patterns can reside together within a narrow range of frequencies, or they can be integrated across frequencies or ears. This view of speech perception, in which across-frequency changes in energy are seen as systematic changes in the temporal fluctuation patterns at two or more fixed loci, is more in line with the physiological encoding of complex signals. [DOI: 10.1121/1.1553464]

PACS numbers: 43.71.Es, 43.71.An [CWT]

I. INTRODUCTION

The production of speech normally involves a vibrating or noisy source which is modified by the resonances of the throat and mouth. Changes in the shape of the resonant cavities cause changes in the spectral distribution of energy, and speech sounds are usually described in these terms. Vowels, for example, are characterized by the position of the lowest three peaks in their spectrum (formant frequencies), which can span several octaves. Consonants can also be distinguished on the basis of their spectral distributions. The fricative sounds distinguishing ''see'' and ''she'' differ primarily in the spectral shape of their stochastic noiselike energy.

However, when spectral information is limited by filtering the speech signal to eliminate some frequencies while passing others, high intelligibility can remain. In a formal study of the intelligibility of speech limited to a narrow ''spectral slit'' (Warren *et al.*, 1995), it was found that narrow-band sentences exhibited very high intelligibilities, despite a lack of features traditionally thought essential for comprehension. When sentences representing everyday speech were restricted to a single $\frac{1}{3}$-octave band having steep filter slopes (96 dB/octave) and center frequencies ranging from 1100 to 2100 Hz, well over 90% of the standard scoring keywords were identified by subjects who were hearing the sentences for the first time and who had no special training. Bands in the surrounding regions down to 530 Hz and up to 4200 Hz produced scores over 60%. These results involving the high intelligibility of narrow-band sentences were confirmed and extended by Stickney and Assmann (2001).

The intelligibility function obtained in the Warren *et al.* study follows the basic pattern of band importance functions which characterize the relative importance of various frequency regions of speech. However, according to the band importance functions of the articulation index (ANSI, 1969/R1986) or speech intelligibility index (ANSI, 1997), individual $\frac{1}{3}$-octave bands in the frequency region that yielded near-perfect intelligibility in the Warren *et al.* study provide only 9% to 11% of the total information in speech. When speech is filtered to a spectral slit, energy is eliminated from the concurrent formants used to distinguish vowels. Similarly, spectral balance, formant transitions, and other spectral features traditionally thought essential for consonant recognition are missing or severely distorted. An obvious feature of the narrow-band sentence is the temporal pattern of amplitude fluctuation, which is characteristic of the particular sentence and of the center frequency of the narrow band.

Recent years have brought a greater appreciation of temporal speech information—cues provided by the pattern of amplitude fluctuation over time (for review, see Cole and Scott, 1974; Van Tasell *et al.*, 1987; Rosen, 1992). The cues present within the temporal waveform (which include envelope, periodicity/voicing, and fine-structure) are thought to be especially important for hearing-impaired listeners, who often possess relatively poor frequency resolution (cf. Moore, 1995; Tyler, 1986), but who appear to process the

[a]Author to whom correspondence should be addressed. Currently at Department of Communication Sciences and Disorders, University of South Carolina, Columbia, SC 29208. Electronic mail: ewh@sc.edu

overall temporal pattern of speech quite well (e.g. Turner *et al.*, 1995).

Much of the work examining the particular aspects of temporal fluctuations that provide linguistic information has focused on very low fluctuation rates (under approximately 20 Hz). Houtgast and Steeneken (1985) used the preservation of these low-frequency fluctuations as a predictor of intelligibility in room acoustics. Drullman *et al.* filtered the temporal fluctuations of speech, limiting them in maximum rate in one study (1994a) and in minimum rate in another (1994b). It was found that the elimination of rates below 4 Hz or above 16 Hz had little effect on speech reception thresholds when all other fluctuations were present, and concluded (1994b) that the temporal modulation spectrum can be divided into equally important parts at 8–10 Hz (but see Drullman *et al.*, 1996). Similar regions of the modulation spectrum have been considered important by Greenberg and his colleagues (Greenberg and Arai, 1998, 2001; Greenberg *et al.*, 1998; Silipo *et al.*, 1999), who proposed the ''modulation spectrogram'' as a technique to visualize the important aspects of speech unobscured by spectro-temporal details (Greenberg and Kingsbury, 1997, also see Kollmeier and Koch, 1994).

Other work has considered the role of faster fluctuations in the perception of speech. Rosen (1992) described that voicing cues are carried by temporal fluctuations of approximately 100–200 Hz, and that fine structure cues exist at rates above that. Although low fluctuation rates can be sufficient for intelligibility, studies employing speech-modulated carrier signals have demonstrated that when spectral information is limited, higher temporal rates (up to 50–200 Hz) can contribute to intelligibility (Van Tasell *et al.*, 1987; Shannon *et al.*, 1995).

Much recent interest in temporal speech information has been generated by research involving cochlear implants. Using an acoustic model of a cochlear implant, Shannon and his colleagues (Shannon *et al.*, 1995; Shannon *et al.*, 1998) reduced spectral information while retaining temporal information, not by narrow-band filtering (as in Warren *et al.*, 1995), but by reducing the spectral detail of broadband speech. The broadband spectrum was partitioned into contiguous frequency regions and the amplitude pattern of each region was used to modulate a corresponding band of noise. Because the temporal pattern of each broad region represented the average of all patterns originally present within the band, individual patterns at individual loci were lost, and only gross spectral information consisting of changes in the relative amplitudes of broad spectral regions remained. Despite these severely reduced spectral cues, high intelligibility of sentences was obtained with as few as three or four separate patterns. Dorman *et al.* (1997) confirmed and extended Shannon *et al.*'s (1995) results using both noise band and tonal carriers. They showed that sentence recognition in quiet reached an asymptote at five channels. For individual vowels, performance asymptoted at eight channels. In an earlier study, Hill *et al.* (1968) found that recognition of individual phonemes by trained listeners leveled off at roughly 75% using five to eight channels.

Due in part to an interest in hearing impairment and prostheses, temporal information in speech has often been studied in conjunction with other usually concurrent cues. A primary nonauditory cue used by impaired listeners is speechreading (lip reading), and a considerable amount of work has shown that temporal fluctuation patterns can aid this cue (e.g., Erber, 1972; Breeuwer and Plomp, 1984, 1986; Grant *et al.*, 1985; Grant *et al.*, 1991, 1994). There have also been other demonstrations that temporal speech patterns can be used in combination with other perceptual cues. Listeners hear noise shaped by the amplitude envelope of broadband words as ''more speechlike'' if the printed version of the word is concurrently available to the listener (Frost *et al.*, 1988; Frost, 1991) or when the shaped noise is presented along with the visual representation of a speaker articulating the words (Repp *et al.*, 1992). In addition to this work, Van Tasell *et al.* (1987) demonstrated some identification of speech based solely on temporal cues. Listeners could identify items from a closed-set of 19 speech syllables (known in advance by the listeners) at above-chance levels based only upon their broadband amplitude envelopes (also see Horii *et al.*, 1971; Van Tasell *et al.*, 1992; Turner *et al.*, 1995).

We know that sentences can retain intelligibility when spectral information is reduced either by filtering to a single narrow band, or by collapsing information from all frequency regions to a limited number of channels. We also know that the fluctuating amplitude pattern of speech contains linguistic information, and that listeners can use this temporal information when combined with other perceptual cues, or when tested on a small closed-set of stimuli. When broadband sentences are filtered to a spectral slit, features long thought essential for comprehension are severely distorted or eliminated. However, intelligibility can remain quite high. An obvious feature of these stimuli is the characteristic temporal pattern of amplitude fluctuation. The following experiments were designed to investigate whether this cue is sufficient to account for the observed near-perfect intelligibility of novel narrow-band sentences.

II. EXPERIMENT 1A: INTELLIGIBILITY OF SINGLE SPEECH-MODULATED TONES

In this experiment, narrow-band sentences were used to modulate the amplitude of sinusoidal tones. The resulting speech-modulated tones followed the overall fluctuating amplitude pattern of the corresponding speech band, but lacked across-frequency differences in this pattern. If the overall temporal pattern of amplitude fluctuation is responsible for the high intelligibility observed with $\frac{1}{3}$-octave sentences, then comparable scores should be obtained when listeners are presented with individual speech-modulated tones.

A. Method

1. Subjects

Thirty listeners (three groups of ten) participated in this experiment. Listeners were recruited from introductory-level psychology courses at the University of Wisconsin—Milwaukee, and received either course credit or money for participating. All were Native English speakers with no known hearing problems and were between the ages of 18 and 40 years (median age = 20 years). Considerable care was

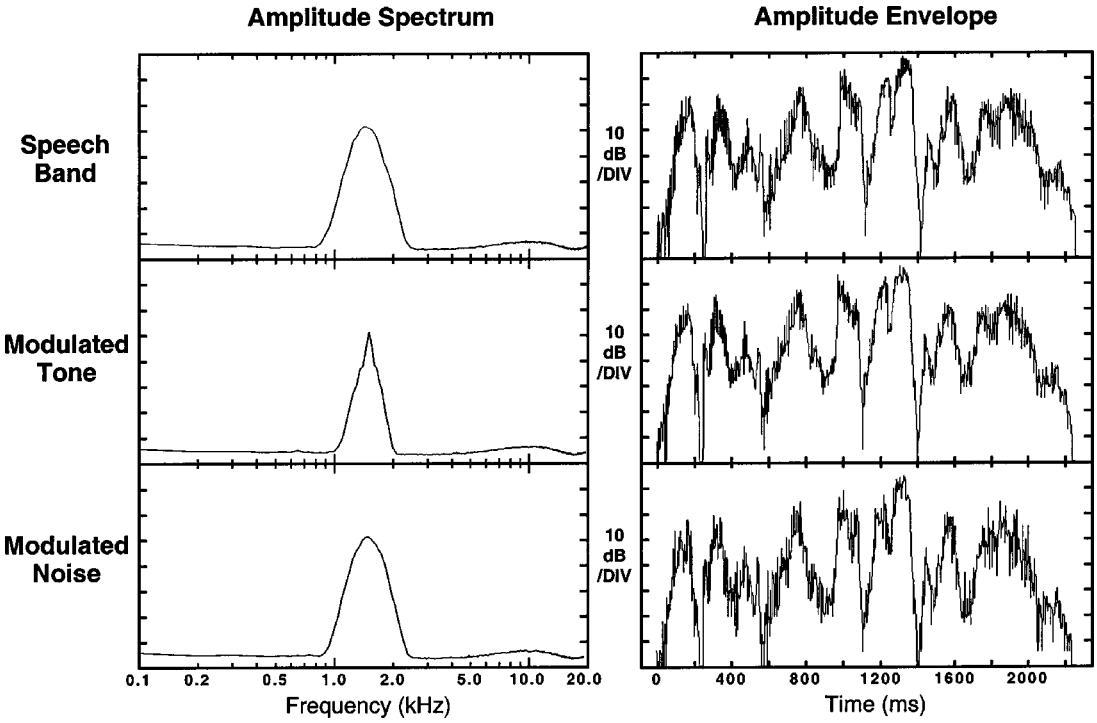

FIG. 1. Similarities between a narrow speech band and narrow speech-modulated bands: The left-most panels show the long-term average amplitude spectrum for a $\frac{1}{3}$-octave band of speech centered at 1500 Hz, a 1500-Hz tone modulated to follow the amplitude pattern of the $\frac{1}{3}$-octave speech (experiment 1a), and a $\frac{1}{3}$-octave band of amplitude-modulated noise (experiment 1b). The panels on the right show the temporal pattern of amplitude fluctuation (rms average of a sliding 2-ms window) traced by each narrow band for the example sentence, ''Walking's my favorite exercise.''

taken to ensure that no listeners had any prior exposure to the sentence materials used in these experiments. Further, participation in any one experimental condition made the subject ineligible for further participation. Thus, all listeners were hearing the stimuli for the first time. These criteria apply to all experiments described here.

2. Stimuli

The stimuli were based upon the Central Institute for the Deaf (CID) ''everyday American speech'' sentences (Silverman and Hirsh, 1955; Davis and Silverman, 1978). The 100 sentences are arranged in ten sets of ten sentences each, with 50 keywords in each set for a total of 500 keywords. An additional set of ten practice sentences was taken from the high-predictability subset of the speech perception in noise (SPIN) test (Kalikow *et al.*, 1977). The sentences were produced using natural rate and intonation by a male speaker having a General American dialect. The recordings were made from within an audiometric chamber (IAC) using a large-diaphram condensor microphone (AKG 414) and digitized at 22 kHz with 16-bit resolution using a Macintosh computer and a Digidesign analog-to-digital converter.

Speech-modulated tones were created at 750, 1500, and 3000 Hz, for comparison with $\frac{1}{3}$-octave speech bands covering a range of center frequencies which provided high intelligibility in the Warren *et al.* (1995) study. The broadband speech file was first filtered to a single $\frac{1}{3}$-octave band using a pair of cascaded eighth-order digital Butterworth filters (yielding slopes of 96 dB/octave). This filtering represents an exact digital emulation of the $\frac{1}{3}$-octave filtering accomplished using analog laboratory filters by Warren *et al.* A speech-modulated tone was created at each center frequency by multiplying a sinusoidal tone having a frequency corresponding to the center of the speech band with the full-wave rectified narrow-band speech on a sample point-by-point basis (cf. Horii *et al.*, 1971). Because the narrow-band filtering of the speech removes the higher temporal fluctuation rates, no low-pass filtering (smoothing) of the rectified speech was employed to ensure that the modulated tones were not limited in temporal rate by some arbitrary low-pass value, and were instead limited only by the narrow-band filtering employed. The modulated patterns were then refiltered to ensure that the modulation sidebands did not exceed the width of the original speech band, and that all energy lied within the original $\frac{1}{3}$-octave region. The same $\frac{1}{3}$-octave (96 dB/octave) filters used to limit the speech were employed to limit the modulated patterns in an identical manner. All processing of the stimuli was performed digitally using MATLAB and Digidesign software. The stimuli were examined at each stage of preparation using a Hewlett–Packard 3561A signal analyzer to ensure that specifications were met.

Each sentence was scaled so that the transduced level (peak of the slow-response rms average) in a flat-plate coupler was within 1 dB of 70 dBA at each ear. The stimuli were converted to analog form and presented to listeners using Digidesign digital-to-analog converters and the Macintosh computer. Analog output was routed to a Mackie (1202-VLZ) mixing board and presented diotically using matched Sennheiser (HD-250) headphones. Figure 1 shows the amplitude spectrum and an example of the temporal pattern for the speech band centered at 1500 Hz, along with those for the corresponding speech-modulated tone employed in the cur-

rent experiment and the speech-modulated noise band employed in experiment 1b.

3. Procedure

Three groups of listeners were randomly assigned to the three modulated tone conditions. Subjects were tested individually, seated across from the experimenter in an audiometric chamber. Listeners were first familiarized with the stimuli and procedures by presenting the 10 SPIN practice sentences. Each listener heard these sentences first broadband, then processed in a manner corresponding to their particular experimental condition. Each listener then heard the 100 CID sentences. They were instructed to repeat each sentence back as accurately as possible during a silent interval separating each sentence. Listeners heard each test sentence only once, received no feedback, and were encouraged to guess if unsure of the content of the sentences. Each listener heard only a single stimulus condition and participated in only a single 30-min session. The experimenter controlled the presentation of each sentence and scored the proportion of keywords reported correctly.

B. Results

A single intelligibility score based on the percentage of 500 keywords accurately reported was calculated for each listener. A group mean intelligibility score and standard error for each condition resulted from averaging these individual means. Panel (a) of Fig. 2 displays the group mean intelligibility scores and standard errors obtained for the three $\frac{1}{3}$-octave speech-modulated tones presented in the current experiment, along with data representing three of the $\frac{1}{3}$-octave speech bands presented to individual groups of subjects by Warren *et al.* (1995). The use of identical speech recordings, identical filtering parameters, the same practice and test procedures, and the same general subject pool allow the results obtained in the current experiment to be directly compared to those obtained previously.[1] In sharp contrast to the high intelligibility of narrow-band sentences, the individual speech-modulated tones exhibited scores near zero. This vast difference between scores was obtained despite the great similarity between the overall temporal fluctuation patterns of the two types of stimuli.

Listener's scores were subjected to a two-factor (2 stimulus types $\times 3$ center frequencies) analysis of variance (ANOVA) which revealed a significant main effect of stimulus type (speech band versus speech-modulated tone) [$F(1,84)=1843.5$, $p<0.0001$], and center frequency [$F(2,84)=35.6$, $p<0.0001$], and a significant interaction [$F(2,84)=38.3$, $p<0.0001$]. The critical comparison is between the speech band and the speech-modulated tone at each center frequency, and individual means comparisons indicated that these scores differed significantly at each frequency [$F(1,28)\geqslant 423.0$, $p<0.0001$]. Simple effects testing of the three narrow speech bands [$F(2,57)=75.9$, $p<0.0001$] revealed that the intelligibility score was different at each center frequency ($p<0.001$ using Scheffé's S). Although the effect of center frequency was also significant for the three modulated tone conditions [$F(2,27)=15.6$, $p<0.0001$], the scores for the 750- and 1500-Hz conditions were equivalent ($p=0.85$), but were significantly lower than that of the 3000-Hz condition ($p<0.001$), which likely accounts for the significant interaction observed in the two-factor ANOVA.

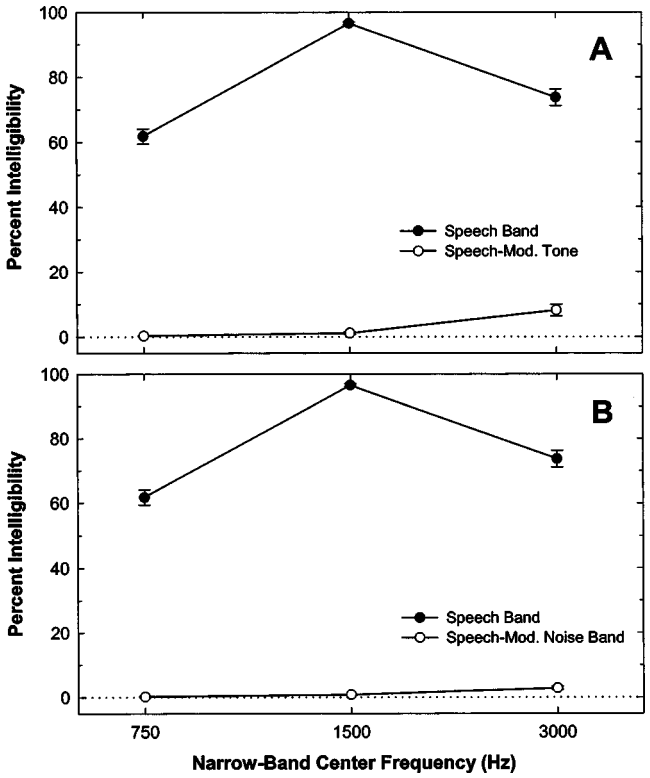

FIG. 2. The high intelligibility of sentences limited to a single $\frac{1}{3}$-octave band, versus the negligible intelligibility of corresponding speech-modulated tones (a) or noise bands (b), at each of three center frequencies: The speech-modulated signals followed faithfully the overall fluctuating amplitude pattern of the narrow speech bands, but lacked contrasting amplitude patterns within the band. Data represent group means and standard errors for 100 sentences containing 500 phonetically balanced keywords heard by each listener. Separate groups of 20 untrained listeners each heard the narrow speech bands and separate groups of ten listeners each heard the speech-modulated stimuli.

III. EXPERIMENT 1B: INTELLIGIBILITY OF SINGLE SPEECH-MODULATED NOISE BANDS

Experiment 1b was an exact replication of experiment 1a, except that speech-modulated noise bands replaced the modulated tones. Like the modulated tones, the $\frac{1}{3}$-octave noise bands followed the overall amplitude fluctuation pattern of the narrow-band speech, but lacked systematic across-frequency differences in this pattern. This replication employing a different type of stimuli and a different modulation technique was performed to ensure that the results obtained with the modulated tones was not a result of any particular manipulation performed.

A. Method

Thirty additional listeners (three groups of ten) were recruited. The speech-modulated stimuli were similar to those in experiment 1a, except that amplitude-modulated noise replaced the amplitude-modulated tone carrier signal at each center frequency (see Fig. 1). Identical speech recordings and prefiltering procedures were used to create $\frac{1}{3}$-octave speech

bands at each of the three center frequencies. The narrow-band speech was converted to speech-correlated noise by randomly reversing the polarity of individual sample points with a probability of 0.5 (Schroeder, 1968). This manipulation maintained the moment-to-moment amplitude (temporal) information within the signal, but destroyed across-frequency differences in the modulation pattern (spectral information) and produced a uniformly amplitude-modulated noise. Because the random reversal of sample points produced a broad speech-modulated noise, the same filters used to create the speech bands were employed to restrict the fluctuating noise pattern in an identical manner. Separate groups of listeners were assigned to the three center frequencies, and experienced procedures identical to those of experiment 1a.

B. Results

Like the modulated tones employed in experiment 1a, the speech-modulated noise bands exhibited near-zero intelligibility scores. Panel (b) of Fig. 2 shows the group mean intelligibility scores and standard errors for the speech-modulated noise bands presented in the current experiment, along with those for the corresponding $\frac{1}{3}$-octave speech bands [data replotted from panel (a)].

A two-factor ANOVA (2 stimulus types $\times 3$ center frequencies) revealed significant main effects of stimulus type (speech band versus speech-modulated noise band) [$F(1,84)=2060.5$, $p<0.0001$], and center frequency [$F(2,84)=37.5$, $p<0.0001$], and a significant interaction [$F(2,84)=37.1$, $p<0.0001$]. As before, the critical comparison is between the speech band and the speech-modulated noise band at each center frequency, and individual means comparisons indicated that these intelligibility scores differed significantly at each frequency [$F(1,28) \geqslant 450.0$, $p<0.0001$]. Simple effects testing revealed the same pattern of results obtained for the speech-modulated tones in experiment 1a—while the scores from the narrow speech bands differed at each of the three center frequencies, the scores from the modulated noise conditions [$F(2,27)=17.7$, $p<0.0001$] were equivalent at 750 and 1500 Hz ($p=0.46$ using Scheffé's S), but were significantly lower than the 3000-Hz condition ($p<0.001$).

A supplementary analysis was conducted to examine differences in scores resulting from the use of speech-modulated tones in experiment 1a, and speech-modulated noise bands in the current experiment. A two-factor ANOVA (2 stimulus types $\times 3$ center frequencies) revealed significant main effects of stimulus type (tone versus noise) [$F(1,54)=8.4$, $p<0.01$], and center frequency [$F(2,54)=24.6$, $p<0.0001$], and a significant interaction [$F(2,54)=6.8$, $p<0.01$]. Individual means comparisons indicated that the group mean scores for the two types of stimuli were effectively equal at 750 Hz and at 1500 Hz [$F(1,18) \leqslant 0.09$, $p \geqslant 0.76$], where all means were at or below 2%, but the score of 8% obtained at 3000 Hz in experiment 1a differed significantly from the score of 3% obtained in experiment 1b [$F(1,18)=22.0$, $p<0.0001$]. The significant interaction was likely due to this relative divergence of scores at 3000 Hz, and equivalence at the other two frequencies.

C. Discussion

The amplitude-modulation techniques employed in experiments 1a and 1b allow the elimination of spectral information from speech by creating a single fluctuating amplitude pattern that corresponds to the average of all those present within a given speech band. When this single amplitude pattern is imposed on a carrier signal, all frequency regions of the carrier follow the pattern and the signal fluctuates homogeneously. When the source signal is broadband speech, the single amplitude-modulated pattern can differ substantially from the source, due to the spread of energy from any one frequency region, to all frequency regions (in the case of modulated noise). It is this spread of energy that destroys spectral characteristics, while maintaining temporal characteristics.

Given this spread of energy and the resulting lack of spectral cues, perhaps it should not be surprising that broadband carrier signals modulated by broadband speech are not intelligible (Frost *et al.*, 1988; Frost, 1991; Shannon *et al.*, 1995), unless listeners are tested on a small closed set of items (e.g., Van Tasell *et al.*, 1987), or shown visual cues to the identity of the speech (e.g., Erber, 1972, 1979). Indeed, due to the effect of peripheral auditory filtering into an ordered series of critical bands, the single amplitude pattern averaged over broadband speech does not correspond to a pattern that would occur at any location on the basilar membrane.

Unlike broadband speech-modulated signals, it can be considered surprising that intelligibility is lost when *narrow-band* speech is converted to a narrow-band speech-modulated signal. Like amplitude-modulation, narrow-band filtering eliminates spectral information. This is accomplished not by averaging information into a single homogeneous pattern, but by eliminating energy which lays outside a narrow spectral slit. The resulting speech band fluctuates in amplitude and traces a temporal pattern which characterizes the particular passage. Similarly, a speech-modulated carrier band derived from this narrow-band speech will encompass a similar narrow band of frequencies and will fluctuate with the temporal pattern of the parent speech band.

It seemed reasonable to speculate that at least some of the high intelligibility observed with narrow-band sentences was based upon temporal information provided by the overall amplitude fluctuation pattern of the spectral slit. [Stickney and Assmann (2001) also recognized the potential contribution of the overall temporal pattern.] After all, if the speech band acts primarily as a single fluctuating pattern, then, unlike broadband signals, there should be little difference between a narrow speech band and a narrow speech-modulated carrier band. However, the current experiments demonstrate that this cue is not sufficient for intelligibility. Rather, experiments 1a and 1b suggest that even within a narrow spectral slit, speech contains spectral information that is essential for its comprehension. Viewed differently, systematic across-frequency changes in energy (spectral information) produce systematic differences in the temporal pattern of amplitude fluctuation at two fixed frequencies. Therefore, the current data suggest that contrasts between different amplitude patterns occur within a narrow speech band, and that these

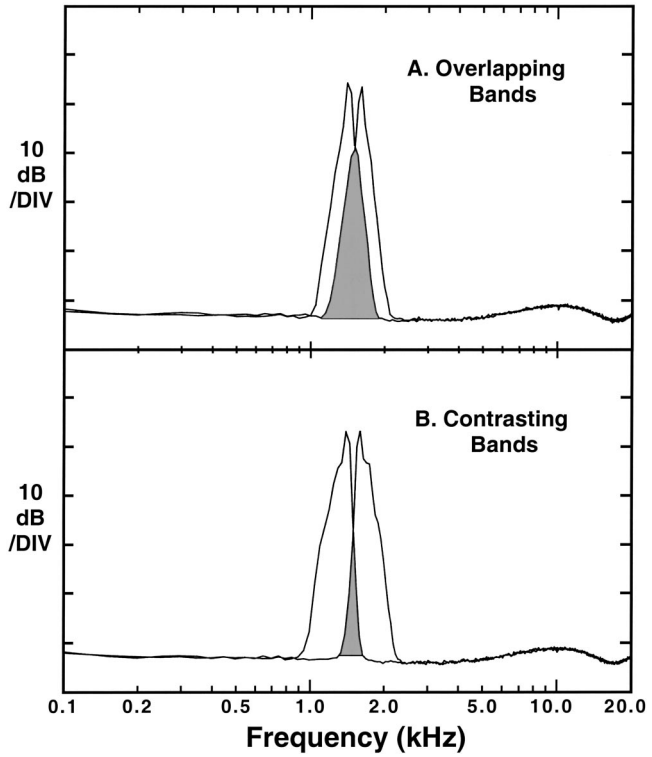

FIG. 3. Dividing a $\frac{1}{3}$-octave speech band into a pair of $\frac{1}{6}$-octave speech-modulated bands: Long-term average amplitude spectra for pairs of separately modulated $\frac{1}{6}$-octave bands (amplitude-modulated tones). The bands shown superimposed in the upper panel overlapped (shaded), disrupting each individual pattern of amplitude fluctuation. In the condition shown in the lower panel, steep filtering (480 dB/octave) was employed along the inner slopes to reduce acoustic overlap and the resultant mixing of adjacent temporal patterns.

contrasting patterns are responsible for the high intelligibility observed with narrow-band sentences.

IV. EXPERIMENT 2A: PARTITIONING A SPEECH BAND INTO A PAIR OF SPEECH-MODULATED TONES

The current experiment is a direct test of the hypothesis that the intelligibility observed with narrow-band sentences is attributable to contrasting patterns of amplitude fluctuation occurring within the narrow band. A $\frac{1}{3}$-octave band of sentences centered at 1500 Hz was divided into a contiguous pair of $\frac{1}{6}$-octaves, and these bands were used to separately modulate the amplitudes of sinusoidal tones. In one of the dual-band conditions, the $\frac{1}{6}$-octave bands overlapped spectrally, potentially disrupting each individual pattern of amplitude fluctuation [see Fig. 3(a)]. The other $\frac{1}{6}$-octave band pair employed very steep filtering to reduce acoustic overlap and to preserve each individual temporal pattern [see Fig. 3(b)]. Each $\frac{1}{6}$-octave speech-modulated band followed the overall amplitude fluctuation pattern of the corresponding speech band, but lacked within-band amplitude contrasts. However, when presented as a pair, they provide a minimal contrast.

A. Method

Forty additional listeners (two groups of 20) were recruited for this experiment. The sentences were first filtered into a pair of contiguous $\frac{1}{6}$-octave bands, corresponding to center frequencies of 1416 and 1589 Hz. The outer slopes for both pairs (the high-pass slopes of the lower bands and the low-pass slopes of the higher bands) were the same as those used to create the $\frac{1}{3}$-octave bands in experiment 1. In the overlapping condition shown in Fig. 3(a), the inner filter slopes (the low-pass slope of the lower band and the high-pass slope of the higher band) were also created using these 96 dB/octave filters and the bands intersected at 1500 Hz, overlapping considerably. In the contrasting dual-band condition shown in Fig. 3(b), the inner slopes were created using ten cascaded passes through eighth-order digital Butterworth filters, producing steep slopes of 480 dB/octave. These bands also intersected at 1500 Hz, but the cascaded filtering produced an extremely-narrow notch between the bands which served to further reduce acoustic overlap. Each $\frac{1}{6}$-octave speech band was used to modulate the amplitude of a sinusoidal tone having a frequency of either 1416 or 1589 Hz, by multiplying the full-wave rectified speech band and the tone on a sample point-by-point basis. The speech-modulated patterns were then postfiltered using the same digital filters used to create the individual $\frac{1}{6}$-octave speech bands. As before, the amplitude of each sentence in each band was adjusted so that its slow rms peak was within 1 dB of the 70 dBA presentation level.

To create the dual-band pairs, the modulated tones were mixed at equal amplitudes using the Digidesign software. To account for the effects of the frequency-dependent phase shifts of the infinite-duration impulse response (IIR) filters, and to ensure proper temporal alignment between members of each pair, the high band in the contrasting pair was delayed relative to the low band by 5 ms. No time correction was required for the overlapping pair. The dual-band pairs were converted to analog form and presented to listeners using the same apparatus employed in experiment 1. Separate groups of listeners were randomly assigned to the two dual-band conditions and heard the 100 CID sentences using procedures identical to those employed in experiment 1.

B. Results

Panel (a) of Fig. 4 shows the group mean intelligibility scores and standard errors for both dual $\frac{1}{6}$-octave band conditions, as well as those for the corresponding $\frac{1}{3}$-octave speech-modulated tone and $\frac{1}{3}$-octave speech band, both from experiment 1a. The intelligibility of the single speech-modulated tone (with its single temporal pattern) was very low, but the narrow band of speech (with its multiple contrasting temporal patterns) produced high intelligibility scores. Some improvement was obtained when the single $\frac{1}{3}$-octave pattern was divided into separately modulated $\frac{1}{6}$-octave bands which overlapped, but the acoustic mixing and resulting disruption of individual temporal patterns in the large region of overlap appears to have hindered intelligibility. However, when the individual modulation patterns of the narrow spectral slits were preserved through steep filtering, an appreciable increase in intelligibility resulted.[2] These scores were subjected to a one-factor ANOVA which revealed a significant effect [$F(3,66)=1101.2$, $p<0.0001$]. Scheffé's *post hoc* tests indicated that all four scores displayed in Fig. 4(a) differ significantly from one another ($p<0.005$).

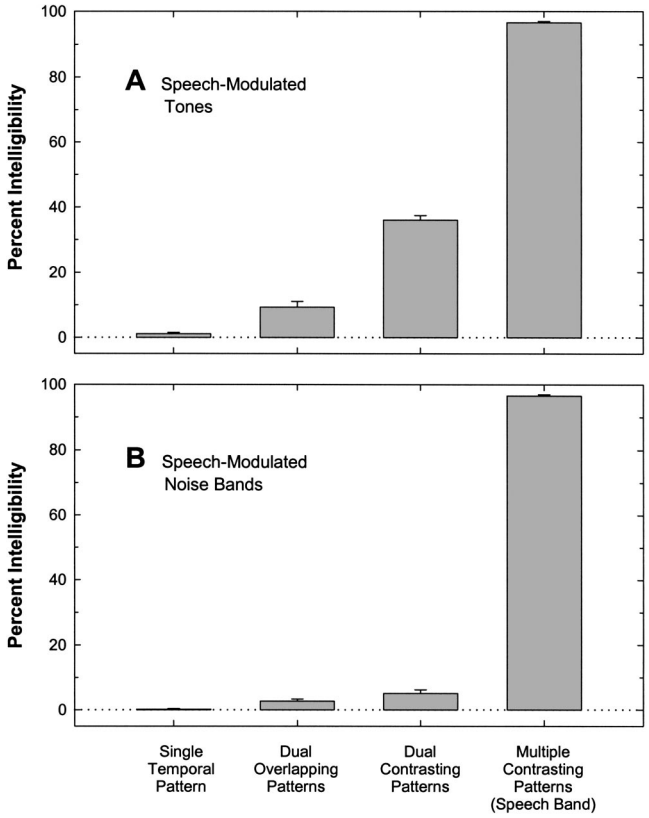

FIG. 4. Intelligibility increases resulting from increasing the amount of contrasting temporal patterns within a single narrow band: Sentences were filtered to a single $\frac{1}{3}$-octave band centered at 1500 Hz (speech band), which was used to create a single $\frac{1}{3}$-octave temporal pattern, or was divided into a pair of separately amplitude-modulated $\frac{1}{6}$-octave patterns. These $\frac{1}{6}$-octave bands either overlapped appreciably [refer to Fig. 3(a)], or had steep filtering along the inner slopes to preserve the individual contrasting patterns [refer to Fig. 3(b)]. Shown are group mean intelligibilities and standard errors for the separate groups of 20 listeners each who heard the dual-band pairs. Scores for the single temporal patterns and for the speech band are replotted from Fig. 2. Each listener heard 100 sentences containing 500 scoring keywords.

V. EXPERIMENT 2B: PARTITIONING A SPEECH BAND INTO A PAIR OF SPEECH-MODULATED NOISE BANDS

As in experiment 1, the effects observed with speech-modulated tones were confirmed using additional groups of listeners, and similar conditions in which speech-modulated noise bands replaced the modulated tones.

A. Method

Both overlapping and contrasting $\frac{1}{6}$-octave band conditions were created and presented to listeners using the same materials, procedures, and apparatus employed in the previous experiment. The same $\frac{1}{6}$-octave filtering procedures were employed to partition the narrow speech band. However, the technique of Schroeder (1968) was employed in place of the tone modulation procedure to create separately modulated noise bands that followed the overall amplitude pattern of the corresponding speech band, but which lacked within-band amplitude contrasts. Postfiltering identical to the $\frac{1}{6}$-octave prefiltering was employed to restrict each modulated noise to its frequency region of origin. Separate groups of 20 listeners each were recruited as before and randomly assigned to the two dual-band conditions.

B. Results

Panel (b) of Fig. 4 shows the group mean intelligibility scores and standard errors obtained for both pairs of $\frac{1}{6}$-octave speech-modulated noise bands, along with those for the corresponding single $\frac{1}{3}$-octave noise band and $\frac{1}{3}$-octave speech band both from experiment 1b. Although considerably lower than the scores produced by modulated tones in experiment 2a, the scores corresponding to these modulated noise bands produced the same statistically significant pattern of results: A one-factor ANOVA revealed a significant effect [$F(3,66) = 3622.0$, $p < 0.0001$], and Scheffé's *post hoc* tests indicated that all four scores differ significantly from one another ($p < 0.05$).

A supplementary analysis was performed to examine differences between the modulated tones employed in experiment 2a and the modulated noise bands employed in experiment 2b. A two-factor ANOVA (2 stimulus types $\times 2$ $\frac{1}{6}$-octave conditions) revealed significant main effects of stimulus type [$F(1,76) = 110.0$, $p < 0.0001$], and $\frac{1}{6}$-octave condition (overlap versus contrast) [$F(1,76) = 194.2$, $p < 0.0001$], and a significant interaction [$F(1,76) = 46.4$, $p < 0.0001$]. Individual means comparisons indicated that the score for the tones was significantly higher than that for the noises in both the overlapping condition [$F(1,38) = 6.8$, $p < 0.05$] and in the contrasting condition [$F(1,38) = 149.6$, $p < 0.0001$], however the difference between scores in the contrasting condition is over four times that of the overlapping condition. The difference in performance resulting from noise band and tonal carriers is perhaps due to random amplitude fluctuations of the modulated noise interfering with details of the temporal speech pattern.

C. Discussion

When narrow-band sentences were replaced with speech-modulated bands in experiment 1, near-perfect intelligibility scores fell to values near zero. Because the two types of stimuli shared similar overall temporal patterns, these results suggested that contrasting temporal patterns exist within the narrow speech bands (but not within the speech-modulated bands) and that these contrasts were responsible for intelligibility. The current experiment supports this conclusion by demonstrating that some intelligibility returned when a single narrow speech band was partitioned into a pair of separately modulated patterns. Thus it appears that contrasting temporal patterns of amplitude fluctuation residing within a narrow band of speech frequencies provide a powerful cue for recognition, and that these contrasting patterns are responsible for the high intelligibility observed with $\frac{1}{3}$-octave sentences.

VI. EXPERIMENT 3A: FREQUENCY SEPARATION OF SPEECH-MODULATED TONE PAIRS

In the following two experiments, pairs of temporal speech patterns having various frequency separations are em-

ployed to examine if the processing of contrasting temporal patterns can be extended to situations in which integration across frequencies is required. Also examined is the enhancement of across-frequency integration resulting from the addition of random noise to the spectral gap between speech-modulated bands. An additional condition was designed to examine dichotic integration of temporal speech information.

A. Method

One-hundred and twenty listeners (six groups of 20) were recruited for this experiment. Dual-band pairs were created by digitally mixing individual 70 dBA $\frac{1}{3}$-octave speech-modulated tones prepared using the procedures of experiment 1a. Five dual-band conditions were prepared, with each band pair having increasing and approximately equal logarithmic separation from 1500 Hz. The band pairs were contiguous (centered at 1336 and 1684 Hz), or were separated by (approximately) one octave (1100 and 2100 Hz), two octaves (750 and 3000 Hz), three octaves (530 and 4200 Hz), or four octaves (370 and 6000 Hz). To correct for the frequency-specific phase shifts of the IIR filters, the high band in each pair was delayed relative to the low band by a value which ranged from 0 to 25 ms and depended upon frequency separation. The delay value required to align each band pair was determined empirically by passing a single-sample click through the digital filters (combining the additive effects of pre- and postfiltering), and measuring the time between the centers of the resulting broad pulses. A dichotic condition was prepared using the modulated tone pair separated by two octaves. Rather than mixing the two bands for diotic presentation, the time-aligned signal was saved to a stereo file so that the two bands could be presented simultaneously to opposite ears. Separate groups of 20 subjects each were randomly assigned to the five dual-band spacing conditions. Each listener heard ten practice sentences and 100 CID test sentences using procedures identical to those employed earlier. The remaining 20 listeners were assigned to the dichotic condition. Ten of those heard the low-frequency band in the right ear and the high-frequency band in the left, and the other ten heard the signals to the two ears reversed.

B. Results

When the $\frac{1}{3}$-octave speech-modulated tones were presented individually in experiment 1a, single-digit intelligibility scores were obtained across all center frequencies tested. However, when the very same patterns were presented in contrasting pairs in the current experiment, appreciable scores were obtained.[3] Figure 5 shows the group mean intelligibility scores and standard errors for the 100 sentences heard by each listener in each dual-band condition. Performance was greatest when the bands were separated by one or two octave(s), and dropped off when the bands were contiguous, or were separated more widely. These scores were subjected to a one-factor ANOVA which revealed a significant effect [$F(4,95)=274.5$, $p<0.0001$]. *Post hoc* analyses using Scheffé's *S* indicated that performance was equivalent for bands separated by one and two octaves ($p=0.18$), and also for bands which were contiguous or separated by three octaves ($p=0.99$). All other comparisons were significant ($p<0.0001$). Also displayed in Fig. 5 is the performance of the dichotic group. Although the individually unintelligible modulated tones produced appreciable intelligibility when presented to opposite ears, the score obtained under dichotic presentation (55%) was lower [$t(38)=7.6$, $p<0.0001$] than that obtained when the same patterns were presented diotically (75%).

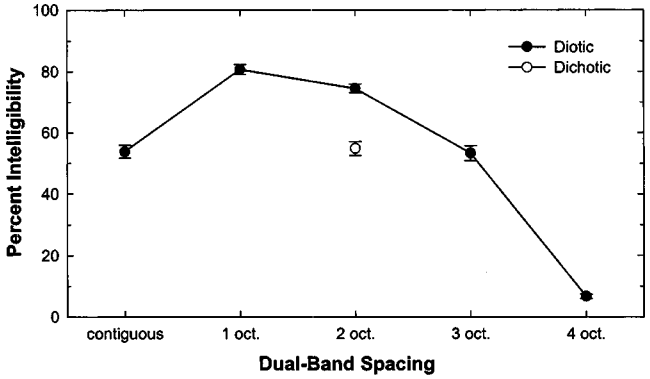

FIG. 5. Single temporal patterns, which are unintelligible when presented individually, can produce appreciable intelligibility when combined into contrasting pairs: Shown are group mean intelligibility scores and standard errors for separate groups of 20 listeners who heard pairs of temporal patterns (tones amplitude-modulated by $\frac{1}{3}$-octave sentences) separated by the values shown. Each listener heard 100 sentences containing 500 keywords. The open symbol represents performance in a dichotic condition in which an additional group of 20 listeners heard the individual temporal patterns presented to opposite ears.

VII. EXPERIMENT 3B: FREQUENCY SEPARATION OF SPEECH-MODULATED NOISE BAND PAIRS

This experiment was similar to experiment 3a, except that speech-modulated noise bands replaced the speech-modulated tones. In an additional set of conditions, a band of Gaussian noise was introduced to the spectral gap between speech-modulated bands to examine the increase in across-frequency integration and intelligibility associated with spectral restoration (Warren *et al.*, 1997).

A. Method

One-hundred and forty additional listeners (seven groups of 20) were recruited. One-third octave speech-modulated noise bands were created using the procedures of experiment 1b. The bands were arranged into five time-aligned pairs using the same frequency separations employed in experiment 3a. Separate groups of 20 listeners each were randomly assigned to the five dual-band pair conditions. The remaining 40 listeners were reserved for the spectral restoration conditions. They were tested using the procedures employed previously.

B. Results

Like the modulated tones, the individually unintelligible modulated noise-bands produced appreciable intelligibility when presented as contrasting pairs. Figure 6 displays the group mean intelligibility scores and standard errors based upon the 500 keywords heard by each listener, for each of the five dual-band conditions.

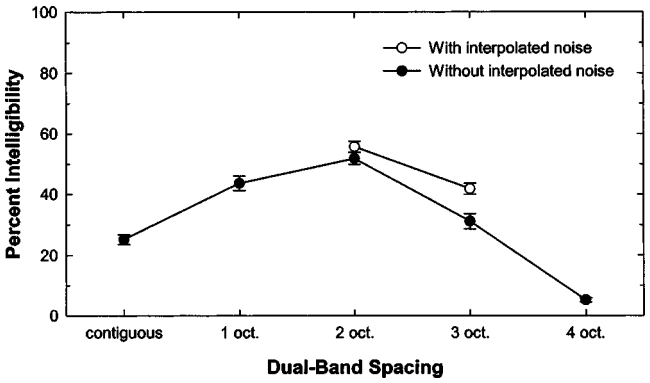

FIG. 6. As in Fig. 5, except that $\frac{1}{3}$-octave speech-modulated noise bands replaced the amplitude-modulated tones. The open symbols represent scores from additional groups of subjects who heard the speech-modulated noise bands along with an interpolated band of Gaussian noise filling the spectral gap between bands.

A one-factor ANOVA [$F(4,95)=83.6$, $p<0.0001$] and subsequent Scheffé tests revealed that, as in experiment 3a, the band pairs separated by one and two octaves produced equivalent scores ($p=0.08$), as did the pairs which were contiguous, or were separated by three octaves ($p=0.34$). All other comparisons were significant ($p<0.001$).

A supplementary analysis was performed to examine differences between the tone and noise band carrier signals. A two-factor ANOVA (2 stimulus types $\times$5 dual-band separations) revealed significant main effects of carrier signal [$F(1,190)=364.2$, $p<0.0001$], and dual band spacing [$F(4,190)=311.4$, $p<0.0001$], and a significant interaction [$F(4,190)=25.0$, $p<0.0001$]. Individual means comparisons indicated that the speech-modulated tone pairs produced higher intelligibility scores than the modulated noise bands at each dual-band separation [$F(1,38)\geqslant 71.6$, $p<0.0001$], except at four octaves where both scores were below 10% [$F(1,38)=0.3$, $p=0.56$].

C. Spectral restoration

Stochastic noise introduced to the spectral gap between speech bands that are widely separated in frequency can produce an increase in intelligibility through spectral restoration (Warren *et al.*, 1997). Like phonemic restoration, spectral restoration allows speech to be intelligible when portions of the signal are subject to masking. In the case of spectral restoration, intelligibility is restored when some speech frequencies are obliterated while others are spared. Although the addition of noise to the spectral gap provides no additional cues to the identity of speech, it is hypothesized to enhance the integration of information which is widely separated in frequency.

Experiments 1 and 2 showed that even very narrow bands of speech can contain contrasting temporal patterns, which account for their high intelligibility. In the current experiment, spectral restoration was examined using speech-modulated bands, rather than speech bands, to eliminate within-band contrasts and strictly isolate across-frequency integration. One-third octave speech-modulated noise bands separated by approximately three octaves (530 and 4200 Hz) and by two octaves (750 and 3000 Hz) were presented along with a continuous band (Wavetek Model 751A, 115 dB/octave) of Gaussian noise (Brüel & Kjær 1405) that filled the spectral gap between bands. To account for the primarily upward spread of masking, the bandwidth of the noise was adjusted so that it just masked sinusoidal tones having frequencies and amplitudes corresponding to the 30-dB down-points of the low-pass skirt of the lower bands and the high-pass skirt of the higher bands. These adjustments resulted in noise bandwidths of 830–2600 Hz for the bands separated by three octaves, and 1150–1750 Hz for the two-octave separation. The level of the noise was set to 60 dBA (rms average), the level found by Warren *et al.* (1997) to provide the maximum increase in intelligibility for speech bands presented at 70 dBA (peak of the slow-response rms average). Two groups of 20 listeners each were assigned to the two conditions. The experimental procedures were identical to those employed previously, except that the continuous interpolated noise was heard along with the speech-modulated bands.

In the previous examination of spectral restoration, noise introduced to the gap separating narrow speech bands having a fixed separation of four octaves produced an increase in intelligibility which averaged 10% across conditions. The group hearing the three-octave separation along with interpolated noise in the current experiment produced a mean intelligibility score of 42%, which is 11% above that produced by the group hearing the otherwise identical condition without noise [$t(38)=3.5$, $p<0.005$]. However, when the band separation was decreased to two octaves, listeners showed a nonsignificant intelligibility increase of 4% over their no-noise counterparts [$t(38)=1.5$, $p<0.16$] (see Fig. 6). The decrease in enhancement as frequency separation was decreased is consistent with the view that spectral restoration reflects an increased tendency to integrate widely separated information that might otherwise be treated as arising from separate sources.

D. Discussion

In experiments 1 and 2, it was shown that subjects possess the ability to effectively process contrasting temporal patterns occurring within a single narrow speech band. The current experiments indicate that these temporal patterns can also be integrated across wide frequency separations.[4] Further, experiments 3a and 3b demonstrate that random noise introduced to the spectral gap between patterns can enhance integration, and that normal-hearing listeners can effectively integrate temporal patterns presented to opposite ears.

VIII. GENERAL DISCUSSION

When broadband speech is filtered to a single narrow band, cues and features long thought essential for comprehension are severely distorted or eliminated. Knowing the rich information content of temporal speech patterns, it seemed reasonable to speculate that the narrow-band pattern of amplitude fluctuation, which remained intact following filtering, could be responsible for the high intelligibility observed with $\frac{1}{3}$-octave sentences. It was found, however, that this cue is not sufficient for intelligibility and that quite a different mechanism is involved. We report here evidence

that contrasting temporal patterns of amplitude fluctuation occurring at different frequency positions within a single narrow speech band are responsible for the high intelligibility of narrow-band sentences. When a narrow band of speech was processed to remove across-frequency spectral information, but maintain its temporal pattern of amplitude fluctuation, intelligibility fell from values near 100% to values near zero. However, when the speech band was divided into a pair of independently amplitude-modulated patterns, some intelligibility returned.

Speech is often characterized by spectral cues produced by systematic across-frequency changes in energy. However, these cues can be viewed differently: Systematic across-frequency changes in speech energy cause the temporal patterns of amplitude fluctuation to differ at different frequency positions. These patterns do not appear sufficient to encode many of the conventional acoustic speech features, especially when the patterns reside together within a narrow spectral slit. They are, however, sufficient for intelligibility of novel sentences by naive listeners, and therefore represent a powerful cue.

In experiment 2, it was found that intelligibility suffered when temporal patterns overlapped spectrally, and these results were attributed to a disruption of individual patterns. Although the mixed pattern will share attributes of its constituents, the contrasting details of the separate patterns are lost in the region of acoustic overlap. These findings are in accord with the widely held view that cochlear implant electrodes need to stimulate discrete populations of neurons to be maximally effective (cf. Wilson *et al.*, 1991). However, Shannon *et al.* (1998) reported a relatively small effect of overlap using a broadband acoustic model of a cochlear implant. To simulate electrode interaction, the authors broadened the filter slopes of the four contiguous speech-modulated noise bands. Performance remained near ceiling levels when slopes were broadened from 24 to 18 dB/octave, but fell when broadened to 6 or 3 dB/octave. Because interactions would have to be quite severe to mimic this substantial overlap, these results imply that speech recognition should be possible for implant patients even with considerable electrode interaction.

Significant differences exist between the stimuli employed by Shannon *et al.* and those employed here. While Shannon *et al.* varied only the presentation carrier bands (in order to most accurately simulate electrode interactions), in the current study, prefiltering of the speech (analysis bands) matched the postfiltering of carriers (presentation bands). However, it may also simply be that the narrow bandwidths and minimum number of channels employed in the current study, and the correspondingly lower intelligibility scores, made these stimuli more sensitive to the hindering effects of overlap. Shannon *et al.*'s standard four-channel broadband stimulus is quite robust, and scores are correspondingly high, and so it may have been better able to resist these effects.

In experiment 3, temporal patterns which were unintelligible when presented individually in experiment 1 were presented in pairs. Because each pattern lacked within-band amplitude contrasts, the contrast necessary for comprehension could only be obtained by integrating across members of the pair. Information was collected regarding the effectiveness of temporal speech information when contiguous and also when widely spaced on the basilar membrane. These conditions allowed an examination of the trade-off between (a) the greater information content of bands from the information-rich center of the spectrum, and (b) the decreased redundancy of widely spaced band pairs. According to the band importance functions of the articulation (ANSI, 1969/1986) and speech intelligibility indexes (ANSI, 1997), speech bands in the region surrounding 1500–2000 Hz provide the greatest relative contribution to intelligibility. However, when a pair of bands is employed, sampling of the speech spectrum is improved if the bands are separated in frequency. The maxima in the functions corresponding to the spacing of band pairs shown in Figs. 5 and 6 seem to reflect this trade-off. Only when bands were separated by four octaves did intelligibility fall to low values. While this may be interpreted as a limit in the ability to integrate information, it is likely that this separation pushes bands into frequency regions having relatively little effective speech information.

Psychophysical studies of envelope correlation have indicated that *same* modulation patterns can be compared across some frequency separation, but that detection of comodulation can decrease as the separation of the patterns increases (e.g., Richards, 1987, 1988; Strickland *et al.*, 1989; Moore and Emmerich, 1990; Takeuchi and Bradia, 1995; also see Richards, 1990). In contrast, the complementary temporal speech patterns employed in the current study produced maximum intelligibility when separated in frequency by one or two octaves. Unlike the randomly or sinusoidally modulated signals typically employed in studies of envelope correlation, temporal speech patterns change systematically as the frequency position of the narrow band changes. In contrast to the across-frequency *comparison* of random modulation patterns, the across-frequency *integration* of complementary temporal speech patterns appears to be quite robust.

As a demonstration of the striking synergy which occurred when individual temporal patterns were combined into contrasting pairs in the current study, the modulated tones, which together produced the highest intelligibility in experiment 3a, were presented individually. One group of ten listeners heard the speech-modulated tone at 1100 Hz, and another group heard the 2100-Hz tone, using materials and procedures identical to those employed earlier. When presented as a contrasting pair, the modulated tones produced a score of 81%. However, when presented individually, the 1100-Hz tone produced a score below 0.5% (standard error of 0.2%) and the 2100-Hz tone produced a score of 0.8% (standard error of 0.4%, see Fig. 7).

Experiments in speechreading have also demonstrated the advantage that pairs of temporal patterns can have over a single pattern. Breeuwer and Plomp (1984) showed that sentence intelligibility scores were increased from 23% for visual information alone, to an average of 49% through the addition of one temporal pattern, and to 75% with two patterns (also see Grant *et al.*, 1991, 1994). While it was reported that pilot testing indicated that the individual narrow-band envelopes were not intelligible when presented without

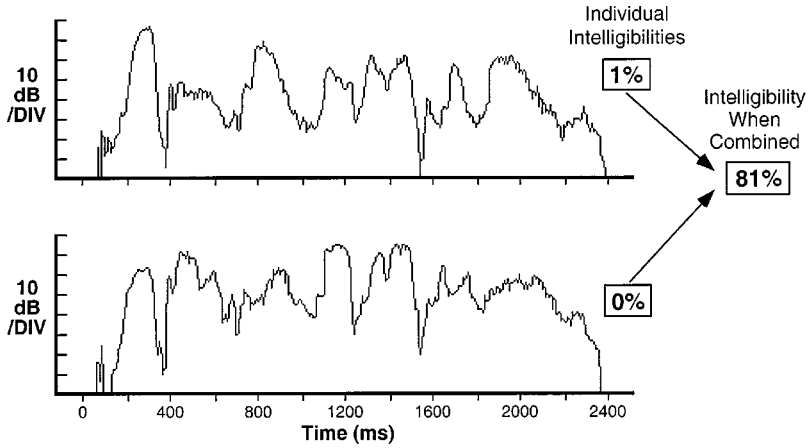

FIG. 7. Synergistic interaction of contrasting temporal speech patterns: The upper panel shows the amplitude tracing (rms average amplitude of a 10-ms sliding window) of a 2100-Hz tone amplitude modulated by a 2100-Hz $\frac{1}{3}$-octave band of speech; the lower panel shows the time-aligned amplitude tracing of an 1100-Hz tone amplitude modulated by an 1100-Hz $\frac{1}{3}$-octave band of speech, both for the sample sentence ''Walking's my favorite exercise.'' Each intelligibility score is based upon separate groups of either 10 listeners (for each single band) or 20 listeners (for the dual-band pair). Each listener heard 100 everyday speech sentences containing 500 phonetically balanced keywords.

visual cues, pairs of these patterns produced an average intelligibility of 20% across conditions. One possible interpretation of the difference between scores obtained by Breeuwer and Plomp and those obtained in the current study involves the 20-Hz low-pass smoothing filter used in the earlier study to limit temporal fluctuation rate.

Throughout these experiments, the scores produced by the speech-modulated noise bands were lower than those obtained using speech-modulated tones. This difference may be attributed to a dilution of temporal speech information due to its mixing with the extraneous amplitude fluctuations of the noise carriers.[5] However, it is noteworthy that the temporal patterns produced by the amplitude-modulated noises were sufficiently close to those of the parent speech bands to allow appreciable intelligibility when presented as contrasting pairs in most conditions. When broad bands of speech are used to create broad amplitude-modulated noises, and when a rectification and multiplication technique is employed in place of the technique of Schroeder (1968), differences between tonal and noise carriers can be quite small (Dorman *et al.*, 1997, also see Shannon *et al.*, 1998). It is possible that the narrow bandwidths employed in the current study accentuated the random amplitude fluctuations of the noise carriers, rendering them less effective than the tones.

In addition to the ability to integrate temporal speech information across frequency separations, experiment 3a showed that listeners can effectively integrate dichotically presented temporal information. These results together indicate that peripheral interactions of temporal information is not necessary, and that these patterns can also combine centrally to produce intelligibility. The integration of information across the two ears is an essential aspect of auditory processing, and disturbances to the auditory system can affect this integration. Tests of binaural fusion (e.g., Matzker, 1959; Linden, 1964; Smith and Resnick, 1972; Willeford, 1977) have typically employed dichotic presentation of complementarily filtered bands of speech. Although the intent is to study across-ear integration, contrasting temporal patterns are present not only *across* the bands at each ear (as intended), but they are also present *within* each speech band. Because the use of speech-modulated bands (rather than speech bands) allows the strict isolation of across-band integration, they may hold promise for the examination of binaural interactions devoid of the interfering effects of contrast processing at each ear.

The current results and the conclusions drawn from them provide a novel framework for the interpretation of speech processing. It is traditional to describe acoustic speech cues according to the distribution of energy in frequency, or the position and movement of energy peaks in the frequency spectrum. Alternatively, speech information has been described according to the temporal information content of single fluctuation patterns from either narrow or broad regions of speech. However, speech is encoded by an array of neural units each responding to pressure changes occurring within a range of frequencies given by the characteristic frequency and response area of the unit. Given this encoding strategy, all information concerning complex dynamic patterns such as speech can be considered contrasting patterns. Rather than assuming that the recognition of speech is accomplished through an extraction and analysis of specific speech features in particular frequency regions, it might be worthwhile to examine correlations between various temporal patterns as speech is produced.

The results presented here provide significant challenges for existing theories of speech perception. Narrow-band speech and narrow-band speech-modulated signals share many characteristics, yet produce dramatically different intelligibility scores. Further, both of these stimuli lack traditional acoustic cues for speech perception. However, these traditional features have largely been established through an examination of individual phonemes and words, rather than speech in its normal context. Due to the robust nature of the speech signal, stimuli used in speech research must be degraded in some fashion so that intelligibility scores fall to values below the ceiling where the effect of various manipulations can be observed. This degradation of the signal has been accomplished in many ways, including filtering and by adding noise. Another traditional method of reducing speech to manageable proportions has been to study isolated phonemes and words.[6] This approach originates from the theoretical belief that the perception of speech primarily involves a ''bottom-up'' analysis of individual phonemes, and a subsequent synthesis into successively larger units. Although most researchers would acknowledge additional ''top-down'' influences provided by the context of speech, this bottom-up

construction of running speech is a basic tenet for much research.

However, speech as a means of communication normally involves the production and perception not of isolated sounds or words, but of sounds strung together into phrases, sentences, and passages that convey a meaningful message. When speech containing some of its everyday context is examined, new cues to its perception emerge. Acoustic cues necessary for the identification of isolated phonemes and words are not required for high intelligibility of everyday sentences. Further, traditional cues revealed in studies examining isolated phonemes and words cannot be considered the only cues available for recognition of speech in normal use, nor can it be safely assumed that these cues are primary in all listening conditions.

IX. SUMMARY AND CONCLUSIONS

The current experiments employed 514 normal-hearing listeners to investigate the role of contrasting temporal patterns in the perception of speech. Experiments 1 and 2 together demonstrate that speech filtered to a narrow spectral slit can retain nearly perfect intelligibility despite a lack of traditional speech features, due to contrasting temporal patterns of amplitude fluctuation within the narrow band. Experiment 3 demonstrated that the use of contrasting temporal patterns is not restricted to situations in which the patterns interact peripherally. Instead, they can be integrated across wide frequency separations, or across the two ears. It is concluded that, despite the rich information content of single temporal patterns, and despite the ability of listeners to use this information to identify speech when tested on a small closed set of items or when presented with additional cues to identity, recognition of novel sentences does not occur. Rather, intelligibility requires a contrast between at least two temporal speech patterns. These patterns can reside together within a narrow speech band, or they can be integrated across wide frequency separations or across the two ears. Speech is typically characterized by acoustic features which change in frequency as the different sounds are produced. However, a position in which across-frequency changes are viewed as changes in amplitude at fixed frequency positions is more in line with the physiological encoding of signals.

ACKNOWLEDGMENTS

These experiments were drawn from a dissertation submitted by the first author to the Graduate School at the University of Wisconsin—Milwaukee. This research was supported in part by NIH\NIDCD Grant No. DC00208 to the second author. Manuscript preparation was supported in part by NIH\NIDCD Grant No. DC01376 to Sid Bacon and by NIH\NIDCD Grant No. DC05795 to the first author. The authors thank Makio Kashino for valuable contributions throughout the course of these experiments, as well as Sid Bacon, Chris Turner, and two anonymous reviewers for thoughtful comments on a previous version of this manuscript.

[1]The high intelligibility of $\frac{1}{3}$-octave (96 dB/octave) sentences observed previously using analog laboratory filtering by Warren *et al.* (1995) was confirmed using two types of digital filters. The first was an exact digital emulation of the previously employed analog filters (dual cascaded passes through eighth-order Butterworth filters). The second was a FIR filter designed to match the amplitude spectrum of the IIR filter (82-order FIR, followed by a 750-Hz high-pass to eliminate low-frequency band-reject ripple). Both were centered at 1500 Hz and implemented in MATLAB. The mean score produced by a group of ten listeners each hearing 100 sentences in the digital IIR condition was 96.2% (standard error of 0.7%), matching the score of 96.6% (standard error of 0.4%) obtained by Warren *et al.* An additional group of ten listeners each hearing 100 sentences in the FIR condition was 98.1% (standard error of 0.4%).

[2]To confirm the results of experiments 1 and 2 under conditions that eliminate the complicating effects of filter skirts and asymmetrical filtering, narrow bands having extremely steep filter slopes were employed. A $\frac{2}{3}$-octave speech band centered at 1500 Hz having filter slopes over 1000 dB/octave (2000-order digital FIR filter) produced a mean intelligibility score of 81.6% (standard error of 1.4%) using a separate group of 20 listeners and the same 100 CID sentences. When that speech band was used to modulate a 1500-Hz tone, a score of only 1.7% was obtained (standard error of 0.4%, $n=10$). However, when the speech band was partitioned into a contiguous pair of $\frac{1}{3}$-octave bands each having extremely steep slopes, and these bands were used to modulate a pair of tones (at 1336 and 1684 Hz), an intermediate score of 31.8% was obtained (standard error of 2.2%, $n=20$) [$F(2,57)=725.8$, $p<0.0001$].

[3]A control condition confirmed the equivalence of the current conditions to those in which a low-pass filter was employed to smooth the temporal fluctuations of the rectified speech. The dual-band pair having a separation of two octaves was recreated with the addition of 200-Hz low-pass filtering (eighth-order Butterworth) following rectification and prior to multiplication with the carrier tones. Four additional subjects heard three lists of ten sentences with low-pass present, alternated with three lists in the original low-pass-absent condition. The group mean intelligibility with low-pass present (83%) did not differ significantly [$t(3)=2.7$, $p>0.05$] from that with low-pass absent (79%). However, when a steeper 200-Hz low-pass filter was employed, scores were reduced somewhat, indicating that temporal fluctuations contained within the filter skirt of the low-pass smoothing filter may potentially contribute to intelligibility.

[4]The 96 dB/octave filter slopes employed in experiment 3 may have caused the modulated patterns to interact at frequency positions having separations smaller than the nominal values. In order to confirm the across-frequency integration of temporal information under conditions in which interactions along the filter slopes were minimal, a pair of $\frac{1}{3}$-octave speech-modulated tones having a separation of two octaves (750 and 3000 Hz) and extremely steep filter slopes (over 1000 dB/oct) was prepared. Modulated tones were prepared as in experiment 3, except that 2000-order digital FIR filters replaced the 96 dB/octave filters in both pre- and postfiltering stages. The contrasting pair of individually unintelligible tones produced a group mean intelligibility score of 38.5% (standard error of 2.8%) using an additional group of 20 listeners, the same 100 sentences, and identical procedures, thus confirming the effective across-frequency integration of temporal information. The scores are lower in this condition when compared to the corresponding 96 dB/octave condition in experiment 3 perhaps because the nominal $\frac{1}{3}$-octave bands have different effective bandwidths due to the presence or absence of filter skirts.

[5]These temporal patterns cannot be accurately considered ''amplitude envelopes'' because they contain not only the slow fluctuations typically associated with that term, but also faster fluctuations capable of carrying (for example) voicing information. Tonal, rather than noise, carriers will better preserve these faster fluctuations.

[6]The intelligibility of narrow-band speech is reduced if isolated words or sentences having low semantic predictability are examined (Stickney and Assmann, 2001).

ANSI (**1969. R 1986**). ANSI-S3.5, 1969 (R 1986), ''American national standard methods for the calculation of the articulation index'' (American National Standards Institute, New York).

ANSI (**1997**). ANSI-S3.5, 1997, ''American national standard methods for the calculation of the speech intelligibility index'' (American National Standards Institute, New York).

Breeuwer, M., and Plomp, R. (**1984**). ''Speechreading supplemented with frequency-selective sound-pressure information,'' J. Acoust. Soc. Am. **76**, 686–691.

Breeuwer, M., and Plomp, R. (**1986**). ''Speechreading supplemented with auditorily presented speech parameters,'' J. Acoust. Soc. Am. **79**, 481–499.
Cole, R. A., and Scott, B. (**1974**). ''Toward a theory of speech perception,'' Psychol. Rev. **81**, 348–374.
Davis, H., and Silverman, S. R. (**1978**). *Hearing and Deafness*, 4th ed. (Holt, Rinehart, and Winston, New York).
Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). ''Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs,'' J. Acoust. Soc. Am. **102**, 2403–2411.
Drullman, R., Festen, J. M., and Plomp, R. (**1994a**). ''Effect of temporal envelope smearing on speech reception,'' J. Acoust. Soc. Am. **95**, 1053–1064.
Drullman, R., Festen, J. M., and Plomp, R. (**1994b**). ''Effect of reducing slow temporal modulations on speech reception,'' J. Acoust. Soc. Am. **95**, 2670–2680.
Drullman, R., Festen, J. M., and Houtgast, T. (**1996**). ''Effect of temporal modulation reduction on spectral contrasts in speech,'' J. Acoust. Soc. Am. **99**, 2358–2364.
Erber, N. P. (**1972**). ''Speech-envelope cues as an acoustic aid to lipreading for profoundly deaf children,'' J. Acoust. Soc. Am. **51**, 1224–1227.
Erber, N. P. (**1979**). ''Speech perception by profoundly hearing-impaired children,'' J. Speech Hear. Disord. **44**, 255–270.
Frost, R. (**1991**). ''Phonetic recoding of print and its effect on the detection of concurrent speech in amplitude-modulated noise,'' Cognition **39**, 195–214.
Frost, R., Repp, B. H., and Katz, L. (**1988**). ''Can speech perception be influenced by simultaneous presentation of print?'' J. Memory Language **27**, 741–755.
Grant, K. W., Braida, L. D., and Renn, R. J. (**1991**). ''Single band amplitude envelope cues as an aid to speechreading,'' Q. J. Exp. Psychol. **43A**, 621–645.
Grant, K. W., Braida, L. D., and Renn, R. J. (**1994**). ''Auditory supplements to speechreading: Combining amplitude envelope cues from different spectral regions of speech,'' J. Acoust. Soc. Am. **95**, 1065–1073.
Grant, K. W., Ardell, L. H., Kuhl, P. K., and Sparks, D. W. (**1985**). ''The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects,'' J. Acoust. Soc. Am. **77**, 671–677.
Greenberg, S., and Arai, T. (**1998**). ''Speech intelligibility is highly tolerant of cross-channel spectral asynchrony,'' Joint Meeting of the ASA and the International Congress on Acoustics, Seattle, pp. 2677–2678.
Greenberg, S., and Arai, T. (**2001**). ''The relation between speech intelligibility and the complex modulation spectrum,'' Eurospeech, Aalborg.
Greenberg, S., and Kingsbury, B. (**1997**). ''The modulation spectrogram: In pursuit of an invariant representation of speech,'' IEEE International Conference on Acoustics, Speech and Signal Processing, Munich, pp. 1647–1650.
Greenberg, S., Arai, T., and Silipo, R. (**1998**). ''Speech intelligibility derived from exceedingly sparse spectral information,'' International Conference on Spoken Language Processing, Sydney, pp. 74–77.
Hill, F. J., McRae, L. P., and McClellan, R. P. (**1968**). ''Speech recognition as a function of channel capacity in a discrete set of channels,'' J. Acoust. Soc. Am. **44**, 13–18.
Horii, Y., House, A. S., and Hughes, G. W. (**1971**). ''A masking noise with speech-envelope characteristics for studying intelligibility,'' J. Acoust. Soc. Am. **49**, 1849–1856.
Houtgast, T., and Steeneken, H. J. M. (**1985**). ''A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria,'' J. Acoust. Soc. Am. **77**, 1069–1077.
Kalikow, D. N., Stevens, K. N., and Elliot, L. L. (**1977**). ''Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability,'' J. Acoust. Soc. Am. **61**, 1337–1351.
Kollmeier, B., and Koch, R. (**1994**). ''Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction,'' J. Acoust. Soc. Am. **95**, 1593–1602.
Linden, A. (**1964**). ''Distorted speech and binaural speech resynthesis test,'' Acta Oto-Laryngol. **58**, 32–48.
Matzker, J. (**1959**). ''Two new methods for the assessment of central auditory functions in cases of brain disease,'' Ann. Otol. Rhinol. Laryngol. **68**, 1185–1197.
Moore, B. C. J. (**1995**). *Perceptual Consequences of Cochlear Damage* (Oxford U.P., Oxford).
Moore, B. C. J., and Emmerich, D. S. (**1990**). ''Monaural envelope correlation perception, revisited: Effects of bandwidth, frequency separation, duration, and relative level of the noise bands,'' J. Acoust. Soc. Am. **87**, 2628–2633.
Repp, B. H., Frost, R., and Zsiga, E. (**1992**). ''Lexical mediation between sight and sound in speechreading,'' Q. J. Exp. Psychol. **45A**, 1–20.
Richards, V. M. (**1987**). ''Monaural envelope correlation perception,'' J. Acoust. Soc. Am. **82**, 1621–1630.
Richards, V. M. (**1988**). ''Components of monaural envelope correlation perception,'' Hear. Res. **35**, 47–58.
Richards, V. M. (**1990**). ''The role of single-channel cues in synchrony perception: The summed waveform,'' J. Acoust. Soc. Am. **88**, 786–795.
Rosen, S. (**1992**). ''Temporal information in speech: Acoustic, auditory and linguistic aspects,'' Philos. Trans. R. Soc. London **336**, 367–373.
Schroeder, M. R. (**1968**). ''Reference signal for signal quality studies,'' J. Acoust. Soc. Am. **44**, 1735–1736.
Shannon, R. V., Zeng, F. G., and Wygonski, J. (**1998**). ''Speech recognition with altered spectral distribution of envelope cues,'' J. Acoust. Soc. Am. **104**, 2467–2476.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). ''Speech recognition with primarily temporal cues,'' Science **270**, 303–304.
Silipo, R., Greenberg, S., and Arai, T. (**1999**). ''Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations,'' Eurospeech, Budapest, pp. 2687–2690.
Silverman, S. R., and Hirsh, I. J. (**1955**). ''Problems related to the use of speech in clinical audiometry,'' Ann. Otol. Rhinol. Laryngol. **64**, 1234–1245.
Smith, B. B., and Resnick, D. M. (**1972**). ''An auditory test for assessing brainstem integrity: Preliminary report,'' Laryngoscope **82**, 414–424.
Stickney, G. S., and Assmann, P. F. (**2001**). ''Acoustic and linguistic factors in the perception of bandpass-filtered speech,'' J. Acoust. Soc. Am. **109**, 1157–1165.
Strickland, E. A., Viemeister, N. F., Fantini, D. A., and Garrison, M. A. (**1989**). ''Within-versus cross-channel mechanisms in detection of envelope phase disparity,'' J. Acoust. Soc. Am. **86**, 2160–2166.
Takeuchi, A. H., and Braida, L. D. (**1995**). ''Effect of frequency transposition on the discrimination of amplitude envelope patterns,'' J. Acoust. Soc. Am. **97**, 453–460.
Turner, C. W., Souza, P. E., and Forget, L. N. (**1995**). ''Use of temporal envelope cues in speech recognition by normal and hearing-impaired listeners,'' J. Acoust. Soc. Am. **97**, 2568–2576.
Tyler, R. S. (**1986**). ''Frequency resolution in hearing-impaired listeners,'' in *Frequency Selectivity in Hearing*, edited by B. C. J. Moore (Academic, London), pp. 309–371.
Van Tasell, D. J., Greenfield, D. G., Logemann, J. J., and Nelson, D. A. (**1992**). ''Temporal cues for consonant recognition: Training, talker generalization, and use in evaluation of cochlear implants,'' J. Acoust. Soc. Am. **92**, 1247–1257.
Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). ''Speech waveform envelope cues for consonant recognition,'' J. Acoust. Soc. Am. **82**, 1152–1161.
Warren, R. M., Riener, K. R., Bashford, J. A., Jr., and Brubaker, B. S. (**1995**). ''Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits,'' Percept. Psychophys. **57**, 175–182.
Warren, R. M., Hainsworth, K., Brubaker, B. S., Bashford, J. A., Jr., and Healy, E. W. (**1997**). ''Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps,'' Percept. Psychophys. **59**, 275–283.
Willeford, J. A. (**1977**). ''Assessing central auditory behavior in children: A test battery approach,'' in *Central Auditory Dysfunction*, edited by R. W. Keith (Grune & Stratton, New York), pp. 43–72.
Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M. (**1991**). ''Better speech recognition with cochlear implants,'' Nature (London) **352**, 236–238.