Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech

Pamela Souza
Department of Speech and Hearing Sciences, University of Washington, 1417 NE 42nd Street, Seattle, Washington 98105

Stuart Rosen
Speech, Hearing and Phonetic Sciences, Division of Psychology and Language Sciences, UCL, 2 Wakefield Street, London WC1N 2PF, United Kingdom

(Received 19 June 2008; revised 16 May 2009; accepted 26 May 2009)

The choice of processing parameters for vocoded signals may have an important effect on the availability of various auditory features. Experiment 1 varied envelope cutoff frequency (30 and 300 Hz), carrier type (sine and noise), and number of bands (2–5) for vocoded speech presented to normal-hearing listeners. Performance was better with a high cutoff for sine-vocoding, with no effect of cutoff for noise-vocoding. With a low cutoff, performance was better for noise-vocoding than for sine-vocoding. With a high cutoff, performance was better for sine-vocoding. Experiment 2 measured perceptibility of cues to voice pitch variations. A noise carrier combined with a high cutoff allowed intonation to be perceived to some degree but performance was best in high-cutoff sine conditions. A low cutoff led to poorest performance, regardless of carrier. Experiment 3 tested the relative contributions of comodulation across bands and spectral density to improved performance with a sine carrier and high cutoff. Comodulation across bands had no effect so it appears that sidebands providing a denser spectrum improved performance. These results indicate that carrier type in combination with envelope cutoff can alter the available cues in vocoded speech, factors which must be considered in interpreting results with vocoded signals.
 [DOI: 10.1121/1.3158835]

PACS number(s): 43.66.Lj, 43.71.Es, 43.71.Bp [RLF] Pages: 792–805

I. INTRODUCTION

Noise- and tone-vocoded signals are used to study the use of temporal cues in normal-hearing (e.g., Shannon *et al.*, 1995; Souza and Turner, 1998; Gallun and Souza, 2008) and hearing-impaired listeners (e.g., Turner *et al.*, 1995; Souza and Boike, 2006) as well as to simulate the information available to cochlear implant users (e.g., Rosen *et al.*, 1999; Shannon *et al.*, 2004; Green *et al.*, 2007). To create these signals, an envelope is extracted from each band of a filter bank and used to modulate a carrier that is either a sine wave at the center frequency of the band or a white noise subsequently filtered to the channel bandwidth. There has been much recent interest in delineating the various factors that are important in accounting for differences between sine- and noise-vocoded speech, and also some areas of disagreement between previous studies. Dorman *et al.*, 1997 is often cited as demonstrating no effect of carrier type for consonants, vowels, or sentences. In contrast, recent studies have found better sentence and/or vowel recognition with sine than with noise carriers (Gonzalez and Oliver, 2005; Chang and Fu, 2006; Whitmal *et al.*, 2007; Stone *et al.*, 2008). For consonants, Whitmal *et al.* (2007) found that sine carriers provided better performance in noise but equivalent performance in quiet. Under some circumstances then, sine-vocoded and noise-vocoded speech do not seem to provide the same information, at least for normal-hearing listeners.

To understand these differences, the authors briefly consider the acoustic form of these signals for a relatively small number of channels, say, 6 or less, where we expect little or no resolution of the harmonics of voiced speech by the analysis filter bank (Fig. 1). Of crucial importance, and interacting with the carrier type, is the cutoff frequency of the envelope smoothing filter that typically follows rectification, as this determines the range of modulation frequencies available in the envelope signal.

A. Vocoding with a sine carrier

When the envelope cutoff is high in comparison to the talker's fundamental frequency (F0), voiced speech will result in envelopes containing amplitude fluctuations corresponding in rate to the voice pitch of the talker. When such an envelope is multiplied against the sinusoidal carrier, sidebands consisting of the sum and difference frequencies of each spectral component in the envelope spectrum and the carrier frequency will be created (as well as a strong component at the carrier frequency due to the strong dc component in the envelope spectrum). Thus the spectrum of each resulting channel output will consist of a series of harmonic-like spectral components centered at the carrier frequency, plus and minus the fundamental frequency. Different from what happens in natural speech, the spectral components squeeze together and expand around their central frequency rather than sweeping all in the same direction during the characteristic F0 glides that constitute intonation. Informal observa-

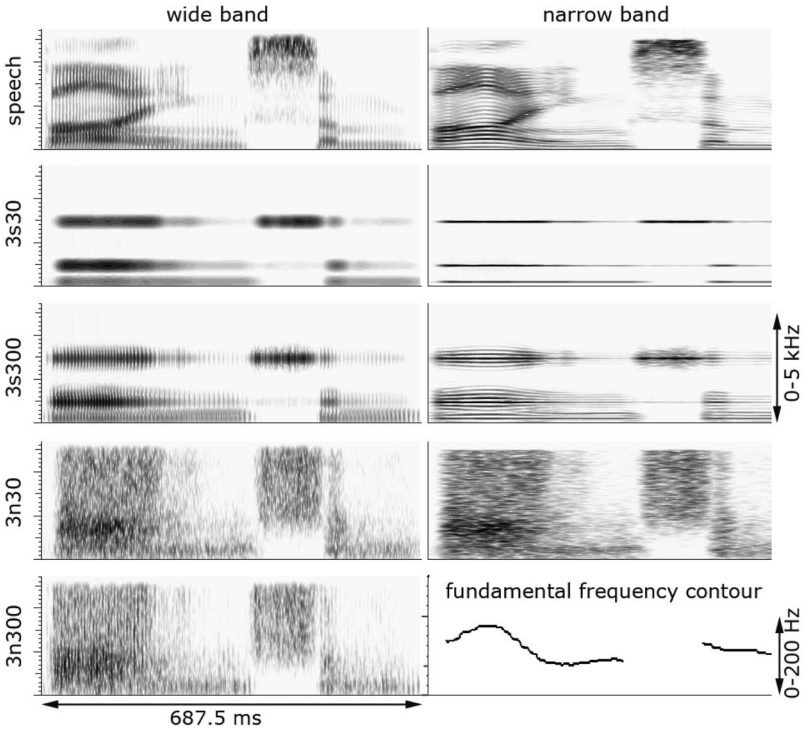

FIG. 1. Wide- and narrow-band spectrograms of various versions of the phrase "buying some" extracted from a male speaker uttering the sentence "They're buying some bread." The top row shows the original speech, followed by examples of the processing used in experiment 1, all based on three-channel sine or noise vocoders with two different envelope smoothing cut-off frequencies (30 and 300 Hz). Hence condition 3s300 refers to a three-channel sine vocoder with an envelope cutoff frequency of 300 Hz. The 30 Hz cutoff was chosen to be significantly below the fundamental frequency (F0) of the talker, and the 300 Hz cutoff chosen to be significantly above. The F0 contour of the utterance can be seen at lower right (narrow-band spectrograms of 3n30 and 3n300 stimuli are visually indistinguishable).

tions suggest that such signals lead to a percept that has a reasonably strong pitch, making information about voicing and intonation accessible to a listener.

Quite apart from the explicitly phonetic information, the common modulation sine components received during voiced speech may also have a beneficial impact on intelligibility. Carrell and Opie (1992) showed that such a common modulation, at least at low enough modulation rates, can improve the intelligibility of sine-wave speech, presumably by making the components cohere better (but note the important acoustic differences between sine-vocoded speech and sine-wave speech).

For aperiodic speech sounds, the range of fluctuation rates in the envelope signal will depend in a straightforward way on the envelope filter cutoff. Higher cutoffs will lead to faster fluctuations, but these will be random in nature. On the one hand, these higher-rate envelope fluctuations may obscure the slower envelope fluctuations. On the other hand, higher-rate envelope fluctuations may cue the presence of aperiodic energy more effectively, a strong cue to voicelessness and fricative manner. However, it is likely that the effect of envelope fluctuations, either for or against aperiodic energy, operates only in the absence of other spectral shape cues (i.e., only in single-channel vocoding) because voiceless excitation tends to have much more energy in high frequency regions than low, the opposite trend to that observed for voiced sounds. Indeed, there is good evidence for the utility of envelope fluctuations above 20 Hz when using a single channel, but not when using higher numbers of channels (Shannon *et al.*, 1995; Fu and Shannon, 2000).

The situation is much simpler when the envelope filter cutoff is low in comparison to F0. Here, any within-channel information about speech periodicity in the envelopes arising from voiced speech will be eliminated and only slow fluctuations will be transmitted. This will result in the output spectrum being dominated by what is essentially a small number of sinusoids varying slowly in amplitude, one for each channel in the vocoder.

B. Vocoding with a noise carrier

When noise is used as a carrier, inherent fluctuations in the noise will be superimposed on the envelope. It seems plausible that these fluctuations (non-existent for a sine carrier) could interfere with or obscure some envelope cues. On the other hand, noise-vocoded consonants result in similar auditory nerve responses to natural speech consonants (Loebach and Wickesberg, 2006) so perhaps the carrier fluctuations are irrelevant.

Varying the envelope filter bandwidth again has its major effects for voiced speech. For envelope filter cutoffs low in comparison to F0, there should be *no* direct cues for periodicity, hence little or no percept of voice pitch. When envelope filter cutoffs are high in comparison to F0, cues to periodicity and intonation are signaled through amplitude modulations in the noise carrier, which may not be very deep, and even in the best situations lead to relatively weak pitches (Burns and Viemeister, 1976, 1981; Patterson *et al.*, 1978). It therefore seems likely that the voice pitch of the talker will be much less salient in noise-vocoded than in sine-vocoded speech at least for high envelope cutoffs. Supporting evidence for this comes from work by Stone *et al.* (2008) who investigated sentence recognition in a single-competing-talker background for six-channel noise and tone carriers. Although this study does not resolve the issue of higher-frequency envelope contributions to speech recognition *per se*, it would be expected that F0 cues would aid talker separation and thus improve performance. In fact, varying envelope bandwidth had much larger effects for sine than noise carriers.

TABLE I. Center and lower-to-upper cutoff frequencies (in hertz) used for the vocoder processing. For each condition the lowest cutoff frequency was 100 Hz and the highest cutoff frequency was 5000 Hz.

	Center frequency (Hz)				
Band	1	2	3	4	5
2	392	2294			
3	269	1005	2984		
4	219	643	1531	3399	
5	192	481	1005	1955	3673
	Lower-to-upper cutoff frequencies (Hz)				
Band	1	2	3	4	5
2	100–1005	1005–5000			
3	100–548	548–1755	1755–5000		
4	100–392	392–1005	1005–2294	2294–5000	
5	100–315	315–705	705–1410	1410–2687	2687–5000

It is also important to note that varying envelope bandwidth should have little impact on the spectrum of noise-vocoded speech. Although the modulation of the envelope by the noise carrier still leads to sidebands, multiplying a white noise by any signal still results in a white noise. Here, the spectral properties of each band are determined by the properties of the filters used to limit the bandwidth of each channel *after* modulation.

In summary, then, envelope cutoff frequency and carrier type, as well as number of bands, interact in a complex way and require further study, especially as regards their effects on intelligibility. Generally speaking, the authors expect much bigger effects of envelope bandwidth for sine-vocoded rather than noise-vocoded speech, because the bandwidth of the extracted envelopes is a crucial determinant of the temporal and spectral characteristics of sine-vocoded speech, but only affects the *temporal* properties of noise-vocoded speech.

Experiment 1 provides a direct assessment of these effects. Two follow-up experiments examine the contributions of fundamental-frequency variations (experiment 2) and carrier comodulation across bands (experiment 3) to the results found.

II. EXPERIMENT 1: ENVELOPE FREQUENCY CUTOFF VERSUS CARRIER TYPE

The purpose of experiment 1 was to vary envelope frequency cutoff, carrier type, and number of bands. The authors expected an interaction between carrier type and envelope cutoff. Previous work also suggested that such an interaction might depend on the type of speech materials.

A. Subjects

Subjects were 16 adults (12 females and 4 males) recruited from the student population at UCL. All were native speakers of southern British English. Subjects ranged in age from 19 to 52 years (mean 25 years). All but one listener had normal hearing, defined as pure-tone thresholds of 20 dB HL or better (see ANSI, 2006) at octave frequencies between 0.25 and 8 kHz. The exception was a single listener (aged 52 years) who had a pure-tone threshold of 30 dB HL at 8 kHz in one ear but who met the 20 dB HL inclusion criteria at all other test frequencies. None of the subjects had any prior experience with the test materials. All subjects were paid for their participation.

B. Stimuli and procedure

All test materials were spoken by the same female talker, who was a native speaker of southern British English. Stimuli were digitally recorded in a quiet room with sampling rates of 22.05 kHz for consonant and vowel stimuli and 48.1 kHz for sentences. The stimuli were vocoded using locally developed MATLAB software, as follows. Each file was digitally filtered into two, three, four, or five bands, using sixth-order Butterworth IIR filters. Filter spacing was based on equal basilar membrane distance (Greenwood, 1990) across a frequency range of 100–5000 Hz. Band center and cutoff frequencies are shown in Table I. Next, the output of each band was half-wave rectified and low-pass filtered (fourth-order Butterworth) at either 30 or 300 Hz to extract the amplitude envelope. The envelope was then multiplied by a carrier, either a tone at the band center frequency or a noise. The resulting signal (envelope $\times$ carrier) was filtered using the same bandpass filter as for the first filtering stage. rms level was adjusted at the output of the filter to match the original analysis, and the signal was summed across bands. Sixteen different conditions were created, each with a unique combination of envelope cutoff frequency (30 or 300 Hz), carrier type (sine or noise), and number of bands (two, three, four, or five).

1. Consonant recognition

Consonant recognition was measured with a set of 20 syllables, including the consonants /b/, /tʃ/, /d/, /f/, /g/, /dʒ/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /ʃ/, /t/, /v/, /w/, /j/, /z/, and /ʒ/ in an /ɑCɑ/ context. These tokens were presented diotically through Sennheiser HD 25 SP headphones at a level between 65 and 70 dB SPL. On each trial, subjects heard a single syllable and selected among a set of orthographic choices using a computer mouse.

To familiarize them with the task, subjects began with a practice block, in which the 20 consonants were sampled from the various test conditions. Each test condition occurred at least once during the practice block. If the response to a trial was correct, visual correct answer feedback was provided and the processed consonant was played again, along with the unprocessed version. If the response was incorrect, the processed version was replayed and the subject was asked to try again. If the second response to that trial was incorrect, the correct answer was shown and the processed and unprocessed consonants were played.

In the test phase, the subject completed one block for each of the 16 test conditions. Each block contained 40 consonants (each token appeared twice). Each subject heard a different order of the conditions based on a Latin square. Correct answer feedback was not provided in the test phase.

2. Vowel recognition

Vowel recognition was measured with a set of ten vowels in a /bVd/ context: "bad" (/æ/), "bard" (/ɑː/), "bawd" (/ɔː/), "bead" (/iː/), "bed" (/e/), "bid" (/ɪ/), "bird" (/ɜː/), "bod" (/ɑ/), "booed" (/uː/), and "bud" (/ʌ/). Testing was similar to that described for consonants except the practice block sampled 30 /bVd/ words from the various test conditions and each test block contained 30 vowels (each token appeared three times).

3. Sentence recognition

Sentences were drawn from the ASL (MacLeod and Summerfield, 1990) and the BKB sentence lists (Bench and Bamford, 1979). Sentence testing was done in an open-set format. The subject repeated the sentence to the experimenter, who was seated in the same room and scored the responses using a computer program, which showed the three key words. The scoring screen was not visible to the subject. Listeners began with a practice block of ten sentences. After responding to a practice sentence, the subject heard the processed and unprocessed versions of that sentence. Each test block consisted of 15 ASL sentences and 16 BKB sentences, with each sentence played once without feedback. No list was repeated. Each sentence had three key words so the score for each condition was based on 93 words.

C. Results

For each reported analysis, Mauchly's (1940) test was evaluated and the Greenhouse–Geisser adjusted values were used if the assumption of sphericity was violated.

1. Consonant recognition

In Fig. 2, proportion correct for each condition is plotted as a function of number of bands. A repeated-measures analysis of variance comparing number of bands, carrier type, and envelope cutoff frequency showed significant main effects of number of bands ($F_{3,45}=275.16$, $p<0.005$), carrier type ($F_{1,15}=10.24$, $p=0.006$), and envelope cutoff frequency ($F_{1,15}=19.54$, $p<0.005$) with no significant three-way interaction ($F_{3,45}=1.60$, $p=0.202$). Carrier type

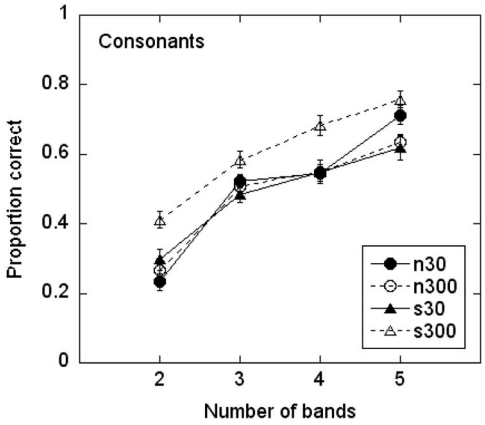

FIG. 2. Proportion correct for consonants as a function of number of bands. Circles show noise carriers; triangles show sine carriers. Filled symbols/solid lines show the 30 Hz envelope cutoff; open symbols/dashed lines show the 300 Hz envelope cutoff. Error bars show ±1 standard error.

interacted with envelope cutoff frequency ($F_{1,15}=23.74$, $p<0.005$) and with number of bands ($F_{3,45}=3.83$, $p=0.016$). Number of bands and envelope cutoff frequency did not interact ($F_{3,45}=1.14$, $p=0.341$).

To understand the nature of the interaction, separate two-way analysis of variance (ANOVAs) (bands × carrier type) were carried out for each envelope cutoff frequency. At a 30 Hz envelope cutoff frequency, performance improved with increasing band number ($F_{3,45}=160.54$, $p<0.005$) and there was no difference between a sine and noise carrier ($F_{1,15}=0.68$, $p=0.423$). A significant bands × carrier interaction [$F_{3,45}=3.84$, $p=0.016$] was caused by better performance ($t_{15}=2.23$, $=0.041$) for the five-band noise carrier, although this difference was no longer significant once Bonferroni-corrected. With a 300 Hz envelope cutoff frequency, performance improved with increasing band number ($F_{3,45}=110.40$, $p<0.005$) and was higher for the sine than for the noise carrier ($F_{1,15}=29.00$, $p<0.005$). The bands × carrier interaction was not significant ($F_{3,45}=1.13$, $p=0.348$).

Information transfer measures were also calculated and are shown in Fig. 3 for voicing (voiced and unvoiced), manner (stop, fricative, nasal, and glide), and place (front, middle, and back). For each condition, the analysis was performed on the composite confusion matrix representing data collapsed across 16 subjects. No place information was available until the signals contained more than two bands, consistent with the fact that place cues rely on spectral distinctions and spectro-temporal dynamics. The sine 300 condition was superior in most respects, particularly for voicing; even the two-band condition transmitted nearly 100% voicing information.

2. Vowel recognition

Results for vowels are shown in Fig. 4. These data were submitted to a repeated-measures analysis of variance (bands × carrier × envelope cutoff). The main effects of bands ($F_{3,39}=144.54$, $p<0.005$) and envelope cutoff frequency ($F_{1,13}=11.00$, $p=0.006$) were significant but the effect of carrier was not ($F_{1,13}=0.401$, $p=0.537$). The three-

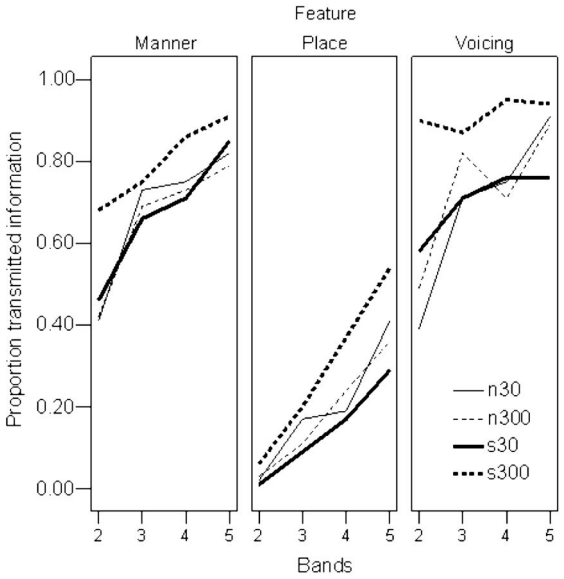

FIG. 3. Transmitted information for voicing, place, and manner for the consonant identification task.

way interaction was nonsignificant ($F_{3,39}=0.80$, $p=0.500$). Unlike the consonants, carrier type and cutoff frequency did not interact ($F_{1,13}=3.03$, $p=0.105$). Instead, the number of bands interacted with both envelope cutoff ($F_{3,39}=5.62$, $p=0.003$) and with carrier type ($F_{3,39}=3.61$, $p=0.021$).

Because bands interacted with both effects of interest, the comparisons of carrier type and envelope cutoff were obtained from two-way ANOVAs (carrier × envelope cutoff), one each for the two-band, three-band, four-band, and five-band stimuli. The carrier × envelope cutoff interaction was nonsignificant ($p>0.05$) in each case. Subjects performed better with the sine carrier than with the noise carrier only for the four-band stimulus ($p=0.001$). For each of the remaining band conditions, scores for the sine and noise carriers were statistically similar. Subjects performed better with the 300 Hz envelope frequency cutoff for four-band ($p=0.009$) and five-band ($p=0.002$) stimuli, but not for the

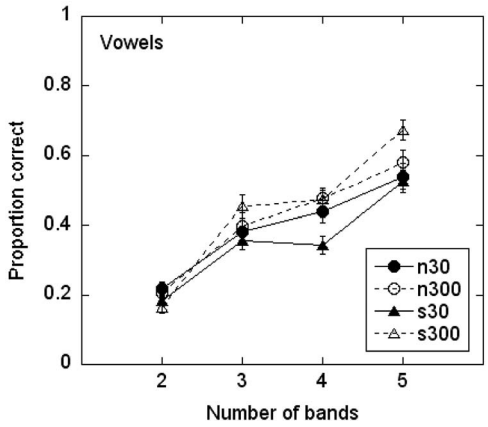

FIG. 4. Proportion correct for vowels as a function of number of bands. Circles show noise carriers; triangles show sine carriers. Filled symbols/solid lines show the 30 Hz envelope cutoff; open symbols/dashed lines show the 300 Hz envelope cutoff. Error bars show ±1 standard error.

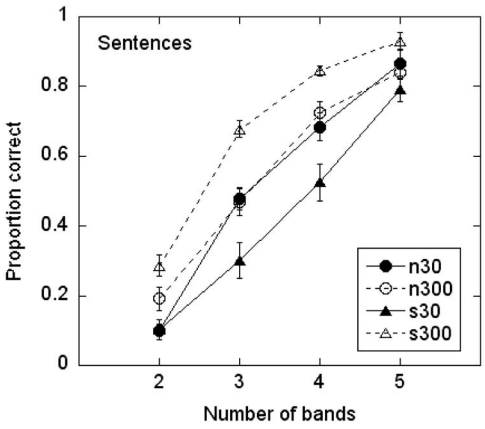

FIG. 5. Proportion correct for sentences as a function of number of bands. Circles show noise carriers; triangles show sine carriers. Filled symbols/solid lines show the 30 Hz envelope cutoff; open symbols/dashed lines show the 300 Hz envelope cutoff. Error bars show ±1 standard error.

two-band or three-band stimuli. This was partially determined by floor effects, as scores for the two-channel stimuli approached chance.

The authors expected that vowels, which are identified by their spectrum, would be better identified as the number of bands increased. In that sense, the lack of improvement between three and four bands for the sine carrier was surprising. To ensure that this finding was not due to any artifact of the vowel token, three subjects repeated the testing with different exemplars from the female talker, as well as a different female talker. As for the original testing, there was little improvement between three and four bands for the sine-vocoded vowel stimuli. A close examination of Dorman *et al.*, 1997 shows a similar effect in which sine-vocoded natural vowels spoken by a single talker show no improvement between four and five bands. This probably occurred because the increase from three to four bands in the current study and from four to five bands in the study of Dorman *et al.* (1997) did not improve representation of the vowel spectra. To understand this, consider a simple example: the vowels /æ/ and /ɑ/ produced by a male talker from the Pacific Northwest. The first formant (F1) for both vowels is approximately 700 Hz, and the second formant (F2) is approximately 1200 Hz for /ɑ/ and 1700 Hz for /æ/ (Bor *et al.*, 2008). Referring to the band cutoff frequencies in Table I, these two vowels would probably not be distinguishable by a three-band vocoder, where they would produce a similar peak in band 2; nor by a four-band vocoder, where they would produce a similar peak in band 3. Only with the five-band vocoder, where F2 for /ɑ/ excites band 3, but F2 for /æ/ excites band 4, would we expect them to be distinguished on the basis of their vocoder response.

3. Sentence recognition

The differences among conditions are more pronounced for sentences (Fig. 5), compared to the vowels and consonants. These data were submitted to a repeated-measures analysis of variance (bands × carrier type × envelope cutoff). The main effects of bands ($F_{3,27}=246.83$, $p<0.005$) and envelope cutoff ($F_{1,9}=34.69$, $p<0.005$) were

significant but that of carrier type was not ($F_{1,9}=0.91$, $p=0.366$). The interaction between carrier type and envelope cutoff frequency ($F_{1,9}=31.36$, $p<0.005$) was significant. The interactions between bands and carrier type and bands ($F_{3,27}=1.08$, $p=0.373$) and between bands and envelope cutoff frequency ($F_{3,27}=2.76$, $p=0.062$) were not significant. The three-way interaction was significant ($F_{3,27}=4.74$, $p=0.009$).

To analyze the various interactions, separate two-way ANOVAs (bands × carrier type) were completed for each envelope cutoff frequency. At a 30 Hz envelope cutoff frequency, scores improved with increasing band number ($F_{3,30}=119.52$, $p<0.005$) and were higher with a noise carrier ($F_{1,10}=10.90$, $p=0.008$). A significant bands × carrier interaction ($F_{3,30}=4.46$, $p=0.010$) was due to a floor effect that precluded any difference between two-band carriers ($p=0.962$), while scores were higher for the noise carrier than the tone carrier with three ($p=0.015$), four ($p=0.002$), and five bands ($p=0.004$). At a 300 Hz envelope cutoff frequency, scores improved with increasing band number ($F_{3,33}=295.51$, $p<0.005$) and were higher with a sine carrier ($F_{1,11}=24.17$, $p<0.005$). The bands × carrier interaction was not significant ($F_{3,33}=1.71$, $p=0.184$).

D. Discussion

Differences among conditions can be summarized as follows: For a 30 Hz envelope cutoff, performance was better for a noise carrier than a sine carrier; for a 300 Hz envelope cutoff, performance was better for a sine carrier than a noise carrier. The authors can also consider results for a single carrier as envelope cutoff frequency increases: This improves recognition for a sine carrier, but not for a noise.

To understand these effects, consider the differences between the various conditions. The first is that with the 300 Hz envelope cutoff in combination with a sine carrier, cues to voice fundamental frequency (voicing and intonation) should be available to the listener (a question the authors explicitly address in experiment 2). Intonation itself has been shown to make a small contribution to sentence intelligibility (Hillenbrand, 2003) but should not matter much for single phoneme contrasts. Temporal information about voicing, although possibly redundant with changes in spectral balance, should help in the identification of consonants but not vowels. From these considerations, the authors would expect manipulating envelope cutoff and carrier type to have its greatest effects for sentences and least for vowels, which, in fact, appears to be the case.

The second major difference between conditions is in the density of the frequency spectrum. As shown in Fig. 1, the 3s30 condition has a sparse spectrum with broad spectral "holes" that may make it more difficult to fuse the percept into a single auditory object. One possibility is that the sidebands in the 300 Hz sine condition and the broader carrier bandwidth of the noise conditions created a denser or more continuous spectrum, which better conveys spectral shape cues. These differences in spectral density are still present after peripheral auditory filtering, as can be seen from the excitation patterns in Fig. 6 (calculated on the basis of normal auditory filtering) to three-channel vocoded versions of a neutral vowel at a fundamental frequency of 150 Hz. Such differences are still apparent for five-channel sine vocoders, although, as here, they are most prominent at lower frequencies where the auditory filters are broadest.

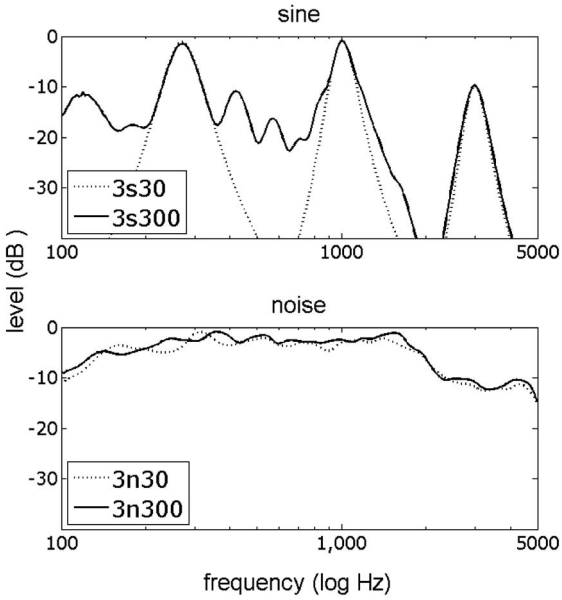

FIG. 6. Excitation patterns to three-channel vocoded versions of a synthetic static neutral vowel (formants at 500, 1500, and 2500 Hz) at a fundamental frequency of 150 Hz. These are calculated on the basis of a gammatone filter bank meant to represent normal auditory filtering (Clark, 2007). Note the lack of difference for a noise-vocoded vowel at the two envelope cutoff frequencies, but the greater degree of difference for the sine-vocoded tokens. Such differences are still apparent for five-channel sine vocoders, although, as here, they are most prominent at lower frequencies where the auditory filters are broadest. For higher numbers of channels, the differences are minor.

We might expect to see this difference in accessibility of spectral shape cues reflected in perception of consonant place, which is carried, primarily, by the frequency spectrum (Rosen, 1992). This was partially supported; place was poorest in the 30 Hz sine condition (Fig. 3) but there was also a difference between the two noise conditions and the 300 Hz sine condition, so a more continuous spectrum cannot be the only explanation.

A third difference is that the 300 Hz envelope cutoff leads to common comodulations across all the carrier bands, which may allow a more ready grouping of those components together. Several studies have shown that common modulations across the individual components of sine-wave speech can improve recognition (Carrell and Opie, 1992; Barker and Cooke, 1999; Lewis and Carrell, 2007), although this may be relatively more important when other cues to auditory object formation are not available. The authors address this issue in experiment 3.

III. EXPERIMENT 2: CORRELATES OF FUNDAMENTAL FREQUENCY

The purpose of experiment 2 was to determine to what extent cues to fundamental frequency (F0) were available in the various test conditions.

A. Subjects

Subjects were ten adults (eight females and two males) recruited from the student populations at UCL (seven subjects) and University of Washington, Seattle (three subjects). Six were native speakers of southern British English, one was a bilingual Swedish/English speaker, and three were native speakers of American English. None had any experience with tonal languages. Subjects ranged in age from 22 to 42 years (mean age 30 years). All had normal hearing, defined as pure-tone thresholds of 20 dB HL or better (see ANSI, 2006) at octave frequencies between 0.25 and 8 kHz. All subjects were paid for their participation.

B. Stimuli and procedure

Two sets of stimuli were used: synthetic glides and naturally produced sentences

1. Synthetic glides

The glide set included four diphthongs: /aʊ/, /eɪ/, /aɪ/, and /oɪ/. Details of stimulus creation are available in Green *et al.*, 2002. These 620-ms long tokens ranged in fundamental frequency such that the F0 at the midpoint in time of each glide was either 113 or 226 Hz. The ratio of start-to-end frequency varied in 12 equal logarithmic steps from 1:0.5 to 1:2.0, so the largest glides went from 80 to 160 Hz or from 160 to 320 Hz. The glides were vocoded using the three-band processing described for experiment 1. Final stimuli included two unprocessed conditions (F0 113 and F0 226) and eight processed conditions (two carrier types $\times$ two envelope cutoffs $\times$ twoF0s).

Stimuli were presented diotically through Sennheiser HD 25 SP headphones at a comfortable listening level. On each trial the subject was required to identify the intonation as either "rise" or "fall" using a computer mouse. A block always consisted of 48 randomly ordered trials (the 4 diphthongs $\times$ 12 F0 steps). The first block in any test condition was intended as a training block and was presented with visual correct answer feedback. The remaining blocks in any test condition were presented without feedback. Subjects completed two blocks in each unprocessed condition, and four blocks in each vocoded condition. The order of test conditions was randomized, with the constraint that the subject completed the unprocessed conditions first.

2. Sentences

Sentences were drawn from 30 sentences previously used by Green *et al.* (2005). These consisted of simple declarative sentences that could be produced as either a statement or question; for example, "They're playing in the garden." The recordings were made in an anechoic room with each of the sentences read as a statement with a falling pitch contour and as a question with a rising pitch contour. One male and one female native talker of Southern British English produced the sentences. The ranges of F0 values were approximately 100–220 Hz for the male talker and 120–360 Hz for the female talker. The sentences were vocoded using the same processing as described for experiment 1. Only three-band versions were created. Final stimuli included five conditions: sine-vocoded with 30 or 300 Hz envelope cutoff, noise-vocoded with 30 or 300 Hz envelope cutoff, and unprocessed; each blocked by gender (male or female talker).

Stimuli were presented diotically through Sennheiser HD 25 SP headphones at a comfortable listening level. On a trial subjects heard a sentence and were required to identify the intonation as either "rise" or "fall" using a computer mouse. A block consisted of 10 trials (for unprocessed sentences) or 20 trials (for vocoded sentences). The first block in any test condition was intended as a training block and was presented with visual correct answer feedback. The remaining blocks in any test condition were presented without feedback. Subjects completed two blocks of the unprocessed speech and four blocks in each vocoded condition. The order of test conditions was randomized, with the restriction that each subject completed the unprocessed conditions first.

C. Results

1. Synthetic glides

Results for the glides are shown in Fig. 7. For the 3s30 and 3n30 conditions at both center F0s, and for the 3n300 condition at the 226 Hz F0, the functions are essentially flat indicating that subjects could not identify the direction of pitch change. A logistic regression was applied to the proportion of fall responses as a function of the log (base 10) of the start-to-end frequency ratio for each processing condition and center F0 for each subject in order to obtain estimates of the slopes of the functions, steeper slopes indicating better performance. Chi squared tests indicated that none of the fits deviated significantly from the observed data.[1] Slope estimates (Table II) ranged from near zero for the relatively flat functions to more than 20 for the unprocessed conditions. The slope estimates for each of the ten subjects were analyzed using a two-way ANOVA. Repeated-measures factors were processing condition and center F0. There was a significant interaction between condition and F0 ($F_{4,60}=4.80$, $p=0.007$), a significant main effect of condition ($F_{4,60}=15.70$, $p=0.001$), but no main effect of F0 ($F_{1,15}=0.35$, $p=0.565$). This is not surprising given that only in condition 3n300 are there obvious differences between the identification functions for the two frequencies. Post-hoc comparisons indicated that results were different for the 113 and 226 Hz F0s for the 3n300 condition ($t_9=2.68$, $p=0.025$) but not for the 3n30 ($t_9=-1.33$, $p=0.217$), 3s30 ($t_9=-0.25$, $p=0.806$), 3s300 ($t_9=-0.45$, $p=0.667$), or unprocessed ($t_7=0.59$, $p=0.576$) conditions.

2. Sentences

Results for the sentences are shown in Fig. 8. A two-way repeated-measures ANOVA (talker $\times$ condition) found no interaction with talker gender, $F_{4,90}=0.550$, $p=0.699$, so data are pooled across talker. There was a significant effect of condition, $F_{4,89}=39.92$, $p<0.005$. All of the conditions were different from one another ($p<0.010$) except for 3n30 and 3n300 ($p=0.081$). Like the glides, performance was best for the 3s300 condition. Unlike the glides, listeners performed above chance in almost all conditions (except 3s30),

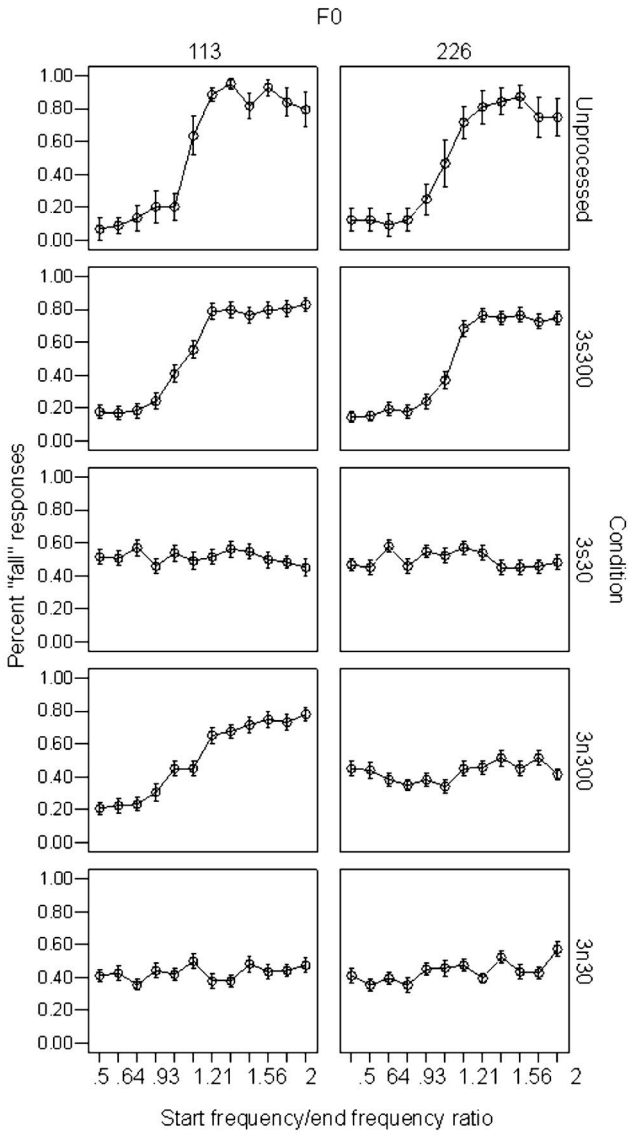

FIG. 7. Percent fall responses in various processing conditions as a function of the start-to-end frequency ratio for synthetic glides. The unprocessed condition is at top.

even for the conditions where no F0 cues to intonation had been available in the glides. Presumably, this reflects availability of other cues to intonation such as word lengthening and envelope cues to syllable (Smith, 2002). Put another way, although F0 derived from temporal envelope cues contributes to sentence intonation, other cues also play a role. In

TABLE II. Mean slope estimate (from a logistic regression) across ten subjects for the proportion of "fall" responses as a function of the start-to-end frequency ratio.

	Center F0	
Condition	113	226
3n30	0.51	0.97
3n300	9.20	0.62
3s30	−0.36	0.02
3s300	9.79	14.34
Unprocessed	22.95	20.92

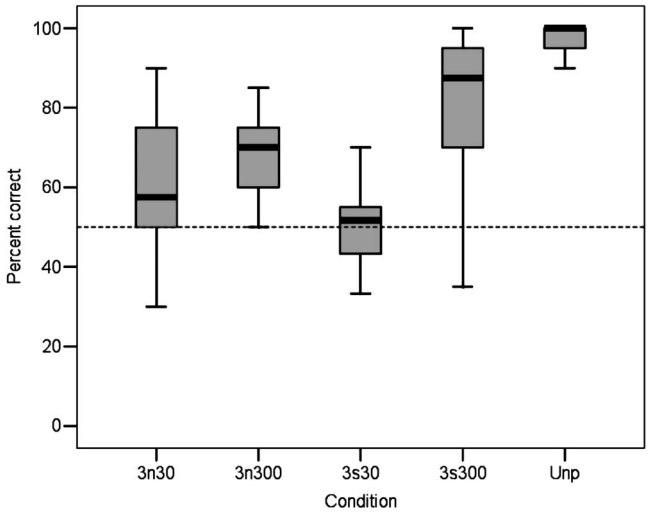

FIG. 8. Boxplot showing the range of scores, interquartile range (box length), and median for each condition in the question/statement task. Chance performance in this two-alternative forced-choice task was 50% (shown by the dashed line).

that sense, the sentence stimuli do not represent a "pure" test of the utility of F0 in judging intonation. However, they do illustrate the partial role of F0 in realistic speech materials. It is interesting that those secondary cues do not seem to be available in the 3s30 condition, where performance was close to chance. One subtle difference between question and statement that can be heard and seen in a spectrogram is the amplitude and duration of the last syllable. It is possible that the 3s30 spectrum is simply too sparse and this cue too weak to be useful.

D. Discussion

Data from experiment 2 demonstrate that F0 information can be derived from envelope fluctuations as long as the envelope cutoff frequency is sufficiently high to allow for transmission of periodicity cues. The effect is largest for the sine carriers, where subjects could not derive any F0 information with a 30 Hz cutoff but performed nearly as well with the 300 Hz cutoff as they did with unprocessed stimuli.

The ability to obtain some pitch change information in the 113 Hz F0, 3n300 but not in the 226 Hz F0, 3n300 condition is consistent with previous work showing that temporal cues to voice pitch are less effective as F0 increases above about 200 Hz (Arehart and Burns, 1999; Green *et al.*, 2002, 2004; Chatterjee and Peng, 2008; Laneau *et al.*, 2006; Patterson *et al.*, 1978). This occurs because sensitivity to modulation decreases with increasing modulation frequency (Grant *et al.*, 1998).

It remains unclear to what extent F0 (versus other cues) contributed to differences in sentence, vowel, and consonant recognition seen in experiment 1. In previous work F0 made only a small contribution to sentence intelligibility (Hillenbrand, 2003) and even less for syllable-length material (Ohde, 1984; Faulkner and Rosen, 1999; Holt *et al.*, 2001). In experiment 1, the authors saw a large improvement in consonant and sentence intelligibility for a 300 Hz versus 30 Hz envelope cutoff. For example, there was a 40% difference

in scores between the 3s30 and 3s300 sentences. The magnitude of the improvement and the similar improvement for consonants as for sentences make it unlikely that F0 is the entire source of the difference between conditions. Accordingly, the authors next examine the contributions of a sparse versus continuous spectrum in combination with across-frequency comodulation.

IV. EXPERIMENT 3: CONTRIBUTIONS OF COMODULATION AND SPECTRAL DENSITY

In experiment 1, performance was best for the 300 Hz sine condition and worst for the 30 Hz sine condition. The authors hypothesized that this resulted because sidebands in the 300 Hz sine condition provided a denser, more continuous spectrum as opposed to the spectral holes prominent in the 30 Hz sine condition. Although the noise carriers would also be expected to exhibit greater spectral density compared to the 30 Hz sine condition, that (theoretical) advantage could be offset by other factors such as the random modulations of the noise carrier itself, or to spectral smearing due to overlap between the noise bands.

Another difference between conditions was the degree to which information was comodulated across bands. Common amplitude modulation may "cohere" the output of separate acoustic filters into a single auditory object. The authors hypothesized that listeners would have more difficulty cohering the sparse information in the 30 Hz sine condition and this might have contributed to performance differences. Because comodulation and spectral shape definition covaried in the original conditions, the purpose of experiment 3 was to create test conditions that varied the degree of comodulation separately from spectral density.

A. Subjects

Subjects were 15 adults (8 females and 7 males) recruited from the student populations at UCL (11 subjects, native speakers of southern British English) and University of Washington, Seattle (4 subjects, native speakers of American English). One subject had participated in experiment 1 and the remaining subjects had no prior experience with the test materials. Subjects ranged in age from 19 to 62 years (mean 30 years). All but two listeners had normal hearing, defined as pure-tone thresholds of 20 dB HL or better (see ANSI, 2006) at octave frequencies between 0.25 and 8 kHz. The two oldest subjects aged 54 and 62 years met the criteria through 4 kHz but had mild (40 dB HL or better) loss at 8 kHz. All subjects were paid for their participation.

B. Stimuli and procedure

Consonant recognition was measured with multiple tokens of the 20 vowel-consonant-vowel utterances used in experiment 1, spoken by a male and a female talker who were native speakers of southern British English. The female talker was the same speaker as for experiment 1. Syllables were produced with the consonants in three different vocalic contexts (/ɑ/, /iː/, and /uː/ hence "ah," "ee," and "oo"). From these, the authors selected six tokens of each combination of vowel and consonant, including three from the male talker and three from the female. Across all exemplars, the mean F0 for the male talker was 91 Hz and the mean F0 for the female talker was 203 Hz.

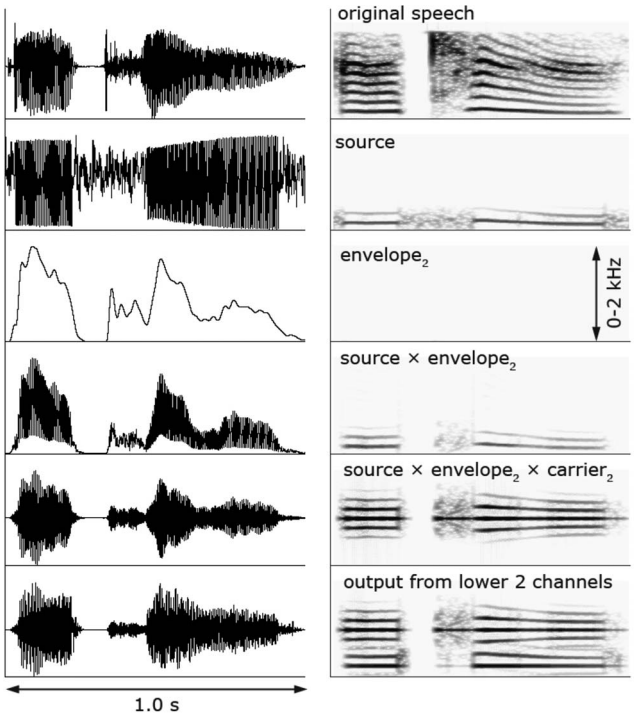

FIG. 9. Time waveforms and narrow-band spectrograms illustrating the various stages in the construction of stimuli in the Px condition of experiment 3. No spectrogram is shown for the envelope signal as it only contains a narrow band of low frequencies. At top is shown the original VCV /ɑ k ɑ/ as spoken by the female talker. Row 2 shows the source function that has a periodic wave matching that of the speech when it is voiced, and a random noise otherwise. Row 3 shows the envelope extracted from channel 2 (hence the subscript), the middle channel of this three-channel vocoding. Row 4 shows the result of multiplying the given envelope by the source, and row 5 the result of a further multiplication by the sine carrier appropriate for this channel. Finally, row 6 shows the final output of channels 1 and 2 summed. The third (highest) channel is not visible because the frequency range of the spectrograms has been limited to 2 kHz so that the harmonics of the voiced source can be clearly observed.

Four test conditions used three-band vocoding with either a sine or noise carrier, and either a 300 or 30 Hz envelope cutoff. These were essentially identical to experiment 1 except that the authors used multiple utterances of the same token, and a male talker as well.

Three further conditions were created to vary comodulation separately from spectral density, all based on three-channel vocoders and a synthetic source. Figure 9 exhibits the various stages in the process. To create a source signal for voiced speech, laryngograph recordings were used to determine the time points of individual vocal fold closures, referred to here as pitch pulses. A speech analysis program (SFS)[2] was used to generate a fundamental-frequency contour based on the pitch pulses, which was used along with the original waveform for hand correction of any pitch pulse errors. The authors then created a sawtooth wave whose periods corresponded to those of the pitch pulses (i.e., which varied in F0 as did the original signal). For each band, the fundamental-frequency modulated sawtooth carrier was filtered to constrain its spectral slope and bandwidth to the

same 300 Hz span that would have occurred for a typically obtained envelope signal using this cutoff. The source for all other intervals of speech was a filtered random noise (i.e., during voiceless speech or silence). See row 2 of Fig. 9 for an example of the source function on its own.

Each speech file was then digitally filtered into three bands, using the same filter bank as previously described. The output of each band was full-wave rectified and low-pass filtered at 30 Hz (fourth-order Butterworth) to extract the amplitude envelope (Fig. 9, row 3 shows the envelope for the middle channel). The source wave was then multiplied by the 30 Hz envelope for each band (Fig. 9, row 4), and the result used to modulate a sine carrier of the appropriate frequency (Fig. 9, row 5). The rms level was adjusted at the output of the filter to match the original analysis and the signal was summed across bands (Fig. 9, row 6). This signal was a control condition termed Px (pulsed excitation), and should be very similar to the 3s300 condition, as can be seen in Fig. 10. One Px condition was created for the male talker and one for the female talker.

A "decoherent" (Dx) signal was created using a constant pitch contour, which varied across channels. For the male speech, the rate was set to 90.7, 79.7, and 110.9 Hz in bands 1–3, respectively, and for the female speech, 202.7, 178.3, and 247.7 Hz in bands 1–3, respectively. The different sawtooth rates were modeled on work by Carrell and Opie (1992) and, as in that study, values were chosen to prevent short-period unintentional comodulation due to a large common factor across the three frequencies.

In order to rule out the possibility that any decrements in performance for the Dx signal resulted from the flattening of the pitch contour as opposed to the differential pulse rates across channels, two sets of monotone signals (Mx) were created using a fixed pulse rate in each channel (91 Hz for the male talker and 203 Hz for the female talker). An example of a stimulus from both the Dx and Mx conditions can be found in Fig. 10.

Test procedures were the same as described for experiment 1 with the following modifications. (1) The practice block consisted of 40 trials, in which the 20 consonants were sampled from the various vowel contexts and test conditions. (2) In the test phase, each block contained 60 trials (20 consonants × 3 vowel contexts). A single block contained stimuli for either the male or female talker. (3) A trial was randomly drawn from the three available exemplars for that talker/consonant/vowel token. (4) The order of test conditions was randomly selected for each subject.

C. Results and discussion

Results, shown in Fig. 11, were analyzed with a two-way (talker × condition) repeated-measures ANOVA. There was no significant effect of talker ($F_{1,28}=2.50$, $p=0.13$), no interaction between talker and test condition ($F_{6,168}=0.78$, $p=0.52$), but a significant effect of test condition ($F_{6,168}=21.79$, $p<0.01$). Post-hoc means comparisons indicated that there was no difference between the Dx and Mx ($t_{29}=0.96$, $p=0.35$), Px and Mx ($t_{29}=-1.40$, $p=0.17$), or Dx and Px ($t_{29}=-0.56$, $p=0.58$) conditions. That is, decohering the spectrum had no significant effect when the density of the spectrum was maintained. This may be thought surprising, since the value of a common (but not independent) modulation of sinusoidal components has been demonstrated previously (Carrell and Opie, 1992). However, in a follow-up study, Lewis and Carrell (2007) also found just as much benefit for independent modulations of the sinusoidal components as for common ones. They noted that the inclusion of sounds such as medial voiceless plosives would cause comodulations across bands (although at lower modulation frequencies), which might have conferred their own cues for auditory grouping, thereby reducing the need for (and benefit of) comodulation due to voice pitch variations. Certainly all of those elements would have been present in our naturally produced speech.

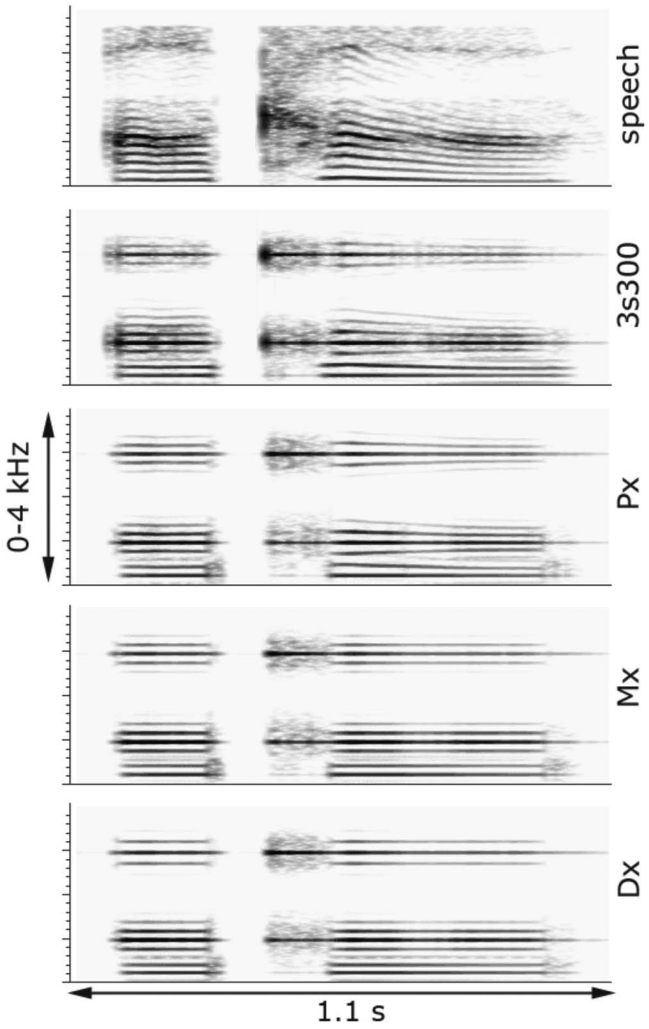

FIG. 10. Narrow-band spectrograms of versions of the VCV /ɑ k ɑ/ as spoken by the female talker of experiment 3. The top panel shows the original speech token, followed by the conditions of particular interest (for details, see the text). Note the similarity between conditions 3s300 and Px. Condition Mx is similar to Px, but with a fixed F0, identical for all three channels, as can be seen from the fact that all the spectral components are fixed in frequency (unlike the varying ones in Px). Condition Dx also uses fixed F0s, but ones which vary from channel to channel (note that the spectral components in the highest frequency channel are more widely spaced than those in the lower two). Also of interest is the fact that the transient release burst of the /k/, shown prominently by the dark "blobs" for both the speech and 3s300 sounds at consonantal onset, is more or less eliminated by the 30 Hz envelope filter used in the other conditions.

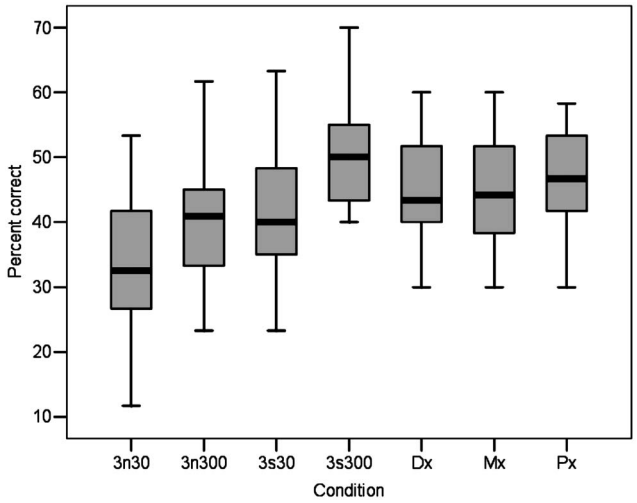

FIG. 11. Boxplot showing the range of scores, interquartile range (box length), and median for each condition in a consonantal identification task.

Although the Px processing was intended to mimic the properties of the 300 Hz sine condition, scores were about 5% higher in 3s300 than in Px ($t_{29}=3.57$, $p<0.01$). The reasons for this are apparent in Table III, which shows confusion matrices for the two conditions collapsed across manner/voicing categories. Performance for voiced fricatives is much better for 3s300, almost certainly attributable to the fact that the Px processing does not allow the representation of mixed excitation (having both a periodic and a noise source) as happens in voiced fricatives such as /ʒ/. Plosives are also better identified in 3s300, probably because the 30 Hz smoothing filter in the Px processing makes the onset burst of the consonant less prominent, as can be seen in the spectrograms in Fig. 9. The only sounds for which Px scores are substantially higher than 3s300 are the nasals but this is balanced out almost exactly by the better performance for glides by 3s300. It is not clear why this difference arises between the two conditions, since the representation of periodicity for such sounds is unlikely to differ much. But glides and nasals are the two classes most often confused in both processing conditions, so may best be thought of as a single category.

TABLE IV. Mean percent correct for experiments 1 and 3, and for only the /ɑ/ vowel context in experiment 3. 16 subjects participated in experiment 1 and 15 (different) subjects participated in experiment 3.

Condition	Experiment 1	Experiment 3 /ɑ/ only	Experiment 3 all vowels
Noise, 30 Hz	52.4	46.0	35.3
Noise, 300 Hz	50.8	54.0	43.3
Sine, 30 Hz	48.3	44.3	41.4
Sine, 300 Hz	58.4	50.7	53.1

As in experiment 1, scores were highest for the 300 Hz sine condition, with lower scores for the 300 and 30 Hz noise conditions and the 30 Hz sine condition. That scores are lower when all vowels were included is not surprising, because more token variability typically leads to lower scores. Table IV compares mean scores for experiment 1, for all vowels in experiment 3, and for only the /ɑ/ vowel context in experiment 3. When only the /ɑ/ vowel is considered, mean scores are generally similar (within 5%) across experiments. The largest difference is the mean score for the 300 Hz sine condition, which is 8% higher in experiment 1. The reasons for this are unclear but in experiment 1 subjects might also

TABLE III. Comparison of confusion matrices collapsed across manner and voicing for the Px (top) and 300 Hz sine (bottom) conditions in experiment 3. For ease of reading, any cell with less than five responses is rendered as blank. "v+" indicates sounds that are voiced and "v−" sounds that are voiceless.

		Response								
Px		v+ plosive	v− affricate	v− fricative	v+ affricate	v− plosive	Glide	Nasal	v+ fricative	Sum
Stimulus	v+ plosive	235			9	17				270
	v− affricate		83			5				90
	v− fricative			255					5	270
	v+ affricate	7			81					90
	v− plosive		40			224				270
	Glide	14					204	60	80	360
	Nasal						46	122	8	180
	v+ fricative	12			10		57		183	270
		Response								
3s300		v+ plosive	v− affricate	v− fricative	v+ affricate	v− plosive	Glide	Nasal	v+ fricative	Sum
Stimulus	v+ plosive	246				13				270
	v− affricate		77		8					90
	v− fricative			253					5	270
	v+ affricate	10			79					90
	v− plosive		26			239				270
	Glide	5					220	58	73	360
	Nasal						66	104	8	180
	v+ fricative	13			20		24		209	270

TABLE V. Mean percent correct for experiments 1 and 3, and for only the /ɑ/ vowel context in experiment 3. Results are for the same five subjects for both experiments.

Condition	Experiment 1	Experiment 3 /ɑ/ only	Experiment 3 all vowels
Noise, 30 Hz	55.6	46.0	31.1
Noise, 300 Hz	51.2	56.0	38.8
Sine, 30 Hz	47.2	44.0	41.8
Sine, 300 Hz	62.9	57.0	48.5

have heard the processed vowels and/or processed sentences before the consonants. It is possible that familiarization affected the more intelligible stimuli to a greater extent. Whitmal *et al.* (2007) also noted the likelihood of such "exposure" effects for their vocoded stimuli.

Other findings replicated the results of experiment 1: For the tone carrier, scores were higher for a 300 Hz than a 30 Hz envelope cutoff ($t_{29}=-5.85$, $p<0.01$); for the 300 Hz cutoff, scores were higher for a tone than for a noise ($t_{29}=-7.13$, $p<0.01$). Those results were verified using a new set of five subjects with normal hearing who completed both experiments, and for whom order of the two experiments was counterbalanced across subjects. Data are shown in Table V and again, show the lowest score for the sine 30 Hz condition when only /ɑ/ vowels were used, and the lowest score for the noise 30 Hz condition when all vowels were included.

V. GENERAL DISCUSSION

Although the differences were most pronounced for the sentences, all materials shared common patterns. With a 30 Hz envelope cutoff frequency, performance was better with a noise than with a sine carrier. With a 300 Hz envelope cutoff frequency, performance was better with a sine than a noise carrier. For sine-vocoded speech, performance was better with a 300 Hz than a 30 Hz envelope cutoff. For noise-vocoded speech, there was no benefit of increasing the envelope cutoff.

To what should these effects be attributed? First, consider the effect of increasing envelope cutoff frequency. According to Rosen's (1992) classification, a 30 Hz envelope cutoff would not transmit cues concerning periodicity (or aperiodicity). Therefore, detection of voicing should be better with the 300 Hz cutoff, and this is supported by the feature analysis.

The higher-frequency envelope modulations would also be expected to provide cues to variations in F0, which can contribute to sentence recognition in quiet and to a smaller extent to consonant recognition in quiet (Faulkner *et al.*, 2000). For example, F0 can code some segmental characteristics such as plosive consonant voicing and aspiration (Haggard *et al.*, 1970). In tonal languages such as Mandarin Chinese, F0 variations are more important than in English (Xu *et al.*, 2002; Xu and Pfingst, 2003). F0 cues will also be important in background noise where they can serve as a cue to talker separation. Stone *et al.* (2008) demonstrated that increasing the envelope cutoff frequency from 45 to 180 Hz improved sentence recognition in a competing speech task even when the signal already offered a high level of spectral detail via a large number of bands. In contrast to the present data, Stone *et al.* (2008) also found some (although small) benefit to increasing envelope filter cutoff for noise-vocoded as well as sine-vocoded signals. This may be related to the nature of the speech-in-noise task as opposed to recognition of speech in quiet; that is, the higher-frequency envelope modulations might have aided talker separation even if not improving recognition *per se*.

Is the advantage of increasing the envelope cutoff frequency with a sine carrier related to the information conveyed by the spectral sidebands, or to the availability of higher-frequency temporal modulations? Gonzalez and Oliver (2005) believed that the ability to identify talker gender using fewer bands with sine-wave carriers than with noise band carriers was partially due to the ability to resolve spectral sidebands for the sine carrier (reflecting the spectral content of the envelope), which of course would signal talker F0, a strong cue to gender. If the effect was solely due to a denser (or more continuous) spectrum, though, the authors would expect to see similar performance for the 300 Hz sine and 300 Hz noise conditions, yet the sine was much superior. Nor can the authors attribute the better performance with the 300 Hz sine condition to a greater sense of comodulation across frequency bands, since "decohering" a signal to abolish across-band asynchrony had no effect on recognition. It therefore appears that the transmission of better cues to periodicity/aperiodicity (signaling voicing) and its variations in F0 (signaling intonation) are the main contributors to this advantage.

Why was it that increasing the envelope cutoff frequency did not confer the same benefit for the noise carrier? Unlike a constant-amplitude sine carrier, a noise carrier has its own envelope fluctuations. These fluctuations could act as interferers, preventing full use of envelope information (Gonzalez and Oliver, 2005), and perhaps negating the benefit of the higher-rate amplitude modulations. Such an effect has already been demonstrated for nonspeech signals (Dau *et al.*, 1999). The idea is that listeners will have more trouble detecting a signal in a randomly-varying masker because the auditory system will not be able to easily distinguish between a signal, and a randomly occurring component of the masker. For whatever reasons, it is well known that both the detection of amplitude modulations in noise carriers and discrimination of modulation rates worsen with increasing modulation rates (Burns and Viemeister, 1976, 1981; Patterson *et al.*, 1978; Viemeister, 1979). A "low-noise noise" can be created by restricting the relationship between noise components such that they will have related phase. Whitmal *et al.* (2007) demonstrated better performance when the signal was vocoded using either sine carriers or low-noise noise carriers, and worse performance with conventional noise carriers. The worst performance was with a narrow-band Gaussian noise carrier, which would have had the greatest inherent (and potentially interfering) envelope fluctuations. A similar

concept might explain the restriction of the benefit of increasing envelope cutoff frequency from 30 to 300 Hz with a noise carrier.

It is also possible that higher-rate temporal fluctuations (i.e., as occur with the higher envelope cutoff frequency) are valuable only when compared to the case where spectral cues alone are insufficient. Xu and co-workers (2005, 2007, 2008) noted a tradeoff such that recognition can be improved either by increasing the envelope cutoff frequency or by increasing the number of spectral bands. In this instance, the 30-Hz noise condition offered about the same level of spectral detail as the 300-Hz noise condition. Perhaps listeners rely more heavily on spectrum to aid identification and the additive effect of higher temporal fluctuations is simply too subtle to contribute in a measurable way.

In summary, our results indicate that changes in carrier type in combination with envelope filter cutoff can significantly alter available cues in vocoded speech. Spectral density and the availability of higher-rate temporal modulations were indicated as major factors. Subjects performed worst for conditions that were spectrally sparse and contained only low-rate temporal modulations. Based on the results of experiment 3, comodulation across bands appeared to make a negligible contribution, at least for these speech materials.

These results also suggest that one should give careful thought to the choice of vocoding method for applications such as cochlear implant simulations. First, consider the level of spectral information provided. In this case, although the number of bands was chosen in part to prevent floor and ceiling effects for our speech tasks, it is within the range of the four to eight effective channels expected in real CI users (Wilson and Dorman, 2008). However, high envelope cut-off frequencies combined with sinusoidal carriers lead to sensitivity to voice pitch variations in normal listeners that far outstrips the performance of any traditional cochlear implant user.

It may also be the case that different simulations mimic different aspects of cochlear implants, but that there is no single implementation that can simulate all aspects and situations. Tone vocoders with high cutoff frequencies are not representative of real implants, but a low envelope cutoff probably results in worse sensitivity to voice pitch than in an actual cochlear implant. Noise-vocoding with high cutoffs probably has appropriate sensitivity to voice pitch, but the fluctuations from the noise itself would not be a factor in a real implant. Although it may not be possible to create a situation where a normal-hearing listener responds as would a cochlear implant wearer in all conceivable ways, different implementations of vocoding can lead to more-or-less realistic results.

ACKNOWLEDGMENTS

The authors thank Tim Green for providing the stimuli for experiment 2, and for many helpful conversations about experimental design and analysis. Many thanks to Sam Eaton-Rosen for crucial help in creating Figs. 1, 9, and 10. The support of NIDCD (Grant No. R01 DC006014) and the Bloedel Hearing Research Foundation to P.S. are gratefully acknowledged, as is the support of the UK Royal National Institute for Deaf People (RNID) to S.R.

[1]Degrees of freedom were either 22 or 46 depending on the number of trials in the block.

[2]Speech Filing System SFS-WIN v 1.5, http://www.phon.ucl.ac.uk/resource/sfs/ (Last viewed 5/14/2009).

ANSI (**2006**). American National Standards Institute S3.22-2006, New York.

Arehart, K. H., and Burns, E. M. (**1999**). "A comparison of monotic and dichotic complex-tone pitch perception in listeners with hearing loss," J. Acoust. Soc. Am. **106**, 993–997.

Barker, J., and Cooke, M. (**1999**). "Is the sine-wave speech cocktail party worth attending?," Speech Commun. **27**, 159–174.

Bench, J., and Bamford, J. (**1979**). *Speech Hearing Tests and the Spoken Language of Hearing Impaired Children* (Academic, London).

Bor, S., Souza, P., and Wright, R. (**2008**). "Multichannel compression: Effects of reduced spectral contrast on vowel identification," J. Speech Lang. Hear. Res. **51**, 1315–1327.

Burns, E. M., and Viemeister, N. F. (**1976**). "Nonspectral pitch," J. Acoust. Soc. Am. **60**, 863–869.

Burns, E. M., and Viemeister, N. F. (**1981**). "Played-again SAM: Further observations on the pitch of amplitude-modulated noise," J. Acoust. Soc. Am. **70**, 1655–1660.

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.

Chang, Y., and Fu, Q.-J. (**2006**). "Effects of talker variability on vowel recognition in cochlear implants," J. Speech Lang. Hear. Res. **49**, 1331–1341.

Chatterjee, M., and Peng, S. C. (**2008**). "Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition," Hear. Res. **235**, 143–156.

Clark, N. (**2007**). GAMMATONE TOOL KIT v2.0 (MathWorks), http://www.mathworks.com/matlabcentral/fileexchange/15313 (Last viewed 5/14/2009).

Dau, T., Verhey, J., and Kohlrausch, A. (**1999**). "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers," J. Acoust. Soc. Am. **106**, 2752–2760.

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.

Faulkner, A., and Rosen, S. (**1999**). "Contributions of temporal encodings of voicing, voicelessness, fundamental frequency, and amplitude variation to audio-visual and auditory speech perception," J. Acoust. Soc. Am. **106**, 2063–2073.

Faulkner, A., Rosen, S., and Smith, C. (**2000**). "Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants," J. Acoust. Soc. Am. **108**, 1877–1887.

Fu, Q. J., and Shannon, R. V. (**2000**). "Effect of stimulation rate on phoneme recognition by nucleus-22 cochlear implant listeners," J. Acoust. Soc. Am. **107**, 589–597.

Gallun, F., and Souza, P. (**2008**). "Exploring the role of the modulation spectrum in phoneme recognition," Ear Hear. **29**, 800–813.

Gonzalez, J., and Oliver, J. C. (**2005**). "Gender and speaker identification as a function of the number of channels in spectrally reduced speech," J. Acoust. Soc. Am. **118**, 461–470.

Grant, K. W., Summers, V., and Leek, M. R. (**1998**). "Modulation rate detection and discrimination by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **104**, 1051–1060.

Green, T., Faulkner, A., and Rosen, S. (**2002**). "Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants," J. Acoust. Soc. Am. **112**, 2155–2164.

Green, T., Faulkner, A., and Rosen, S. (**2004**). "Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants," J. Acoust. Soc. Am. **116**, 2298–2310.

Green, T., Faulkner, A., Rosen, S., and Macherey, O. (**2005**). "Enhancement of temporal periodicity cues in cochlear implants: Effects on prosodic perception and vowel identification," J. Acoust. Soc. Am. **118**, 375–385.

Green, T., Katiri, S., Faulkner, A., and Rosen, S. (**2007**). "Talker intelligibility differences in cochlear implant listeners," J. Acoust. Soc. Am. **121**,

EL223–229.
Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.
Haggard, M., Ambler, S., and Callow, M. (**1970**). "Pitch as a voicing cue," J. Acoust. Soc. Am. **47**, 613–617.
Hillenbrand, J. (**2003**). "Some effects of intonation contour on sentence intelligibility," J. Acoust. Soc. Am. **114**, 2338.
Holt, L. L., Lotto, A. J., and Kluender, K. R. (**2001**). "Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement?," J. Acoust. Soc. Am. **109**, 764–774.
Laneau, J., Moonen, M., and Wouters, J. (**2006**). "Factors affecting the use of noise-band vocoders as acoustic models for pitch perception in cochlear implants," J. Acoust. Soc. Am. **119**, 491–506.
Lewis, D. E., and Carrell, T. D. (**2007**). "The effect of amplitude modulation on intelligibility of time-varying sinusoidal speech in children and adults," Percept. Psychophys. **69**, 1140–1151.
Loebach, J. L., and Wickesberg, R. E. (**2006**). "The representation of noise vocoded speech in the auditory nerve of the chinchilla: Physiological correlates of the perception of spectrally reduced speech," Hear. Res. **213**, 130–144.
MacLeod, A., and Summerfield, Q. (**1990**). "A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use," Br. J. Audiol. **24**, 29–43.
Mauchly, J. W. (**1940**). "Significance test for sphericity of a normal n-variate distribution," Ann. Math. Stat. **11**, 204–209.
Ohde, R. N. (**1984**). "Fundamental frequency as an acoustic correlate of stop consonant voicing," J. Acoust. Soc. Am. **75**, 224–230.
Patterson, R. D., Johnson-Davies, D., and Milroy, R. (**1978**). "Amplitude-modulated noise: The detection of modulation versus the detection of modulation rate," J. Acoust. Soc. Am. **63**, 1904–1911.
Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London **336**, 367–373.
Rosen, S., Faulkner, A., and Wilkinson, L. (**1999**). "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," J. Acoust. Soc. Am. **106**, 3629–3636.
Shannon, R. V., Fu, Q. J., and Galvin, J., III (**2004**). "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation," Acta Oto-Laryngol. **552**, 50–54.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Smith, C. L. (**2002**). "Prosodic finality and sentence type in French," Lang Speech **45**, 141–178.
Souza, P., and Boike, K. T. (**2006**). "Combining temporal-envelope cues across channels: Effects of age and hearing loss," J. Speech Lang. Hear. Res. **49**, 138–149.
Souza, P., and Turner, C. W. (**1998**). "Multichannel compression, temporal cues and audibility," J. Speech Lang. Hear. Res. **41**, 315–326.
Stone, M. A., Fullgrabe, C., and Moore, B. C. J. (**2008**). "Benefit of high-rate envelope cues in vocoder processing: Effect of number of channels and spectral region," J. Acoust. Soc. Am. **124**, 2272–2282.
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.
Viemeister, N. F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. **66**, 1364–1380.
Whitmal, N. A., Poissant, S. F., Freyman, R. L., and Helfer, K. S. (**2007**). "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," J. Acoust. Soc. Am. **122**, 2376–2388.
Wilson, B. S., and Dorman, M. F. (**2008**). "Cochlear implants: A remarkable past and a brilliant future," Hear. Res. **242**, 3–21.
Xu, L., and Pfingst, B. E. (**2003**). "Relative importance of temporal envelope and fine structure in lexical-tone perception (L)," J. Acoust. Soc. Am. **114**, 3024–3027.
Xu, L., and Pfingst, B. E. (**2008**). "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," Hear. Res. **242**, 132–140.
Xu, L., Thompson, C. S., and Pfingst, B. E. (**2005**). "Relative contributions of spectral and temporal cues for phoneme recognition," J. Acoust. Soc. Am. **117**, 3255–3267.
Xu, L., Tsai, Y., and Pfingst, B. E. (**2002**). "Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses," J. Acoust. Soc. Am. **112**, 247–258.
Xu, L., and Zheng, Y. (**2007**). "Spectral and temporal cues for phoneme recognition in noise," J. Acoust. Soc. Am. **122**, 1758–1764.