Differential contribution of envelope fluctuations across frequency to consonant identification in quiet

Frédéric Apoux^{a)} and Sid P. Bacon

*Psychoacoustics Laboratory, Department of Speech and Hearing Science, Arizona State University,
P.O. Box 870102, Tempe, Arizona 85287-0102*

(Received 3 November 2006; revised 22 January 2008; accepted 25 February 2008)

Two experiments investigated the effects of critical bandwidth and frequency region on the use of temporal envelope cues for speech. In both experiments, spectral details were reduced using vocoder processing. In experiment 1, consonant identification scores were measured in a condition for which the cutoff frequency of the envelope extractor was half the critical bandwidth (HCB) of the auditory filters centered on each analysis band. Results showed that performance is similar to those obtained in conditions for which the envelope cutoff was set to 160 Hz or above. Experiment 2 evaluated the impact of setting the cutoff frequency of the envelope extractor to values of 4, 8, and 16 Hz or to HCB in one or two contiguous bands for an eight-band vocoder. The cutoff was set to 16 Hz for all the other bands. Overall, consonant identification was not affected by removing envelope fluctuations above 4 Hz in the low- and high-frequency bands. In contrast, speech intelligibility decreased as the cutoff frequency was decreased in the midfrequency region from 16 to 4 Hz. The behavioral results were fairly consistent with a physical analysis of the stimuli, suggesting that clearly measurable envelope fluctuations cannot be attenuated without affecting speech intelligibility. © 2008 Acoustical Society of America. [DOI: 10.1121/1.2897916]

PACS number(s): 43.71.An, 43.71.Es, 43.66.Ba [KWG]

Pages: 2792–2800

I. INTRODUCTION

For continuous speech in quiet, the dominant components of the modulation spectrum—the spectral representation of the temporal envelope—lie between 1 and 16 Hz with a peak around 4 Hz, reflecting the average syllable rate in speech (Arai and Greenberg, 1997; Greenberg, 1999). The speech modulation spectrum, however, also contains amplitude components above 16 Hz (Plomp, 1983; Houtgast and Steeneken, 1985). Consistent with the physical characteristics of the temporal structure of speech, a number of behavioral studies have demonstrated that speech intelligibility is only affected by reducing or removing modulation frequencies below a few tens of hertz, suggesting that higher envelope fluctuations do not contribute to overall intelligibility (Drullman *et al.*, 1994; Shannon *et al.*, 1995; Arai *et al.*, 1999; Xu *et al.*, 2005).

Although both physical and behavioral data suggest a limited role of high-frequency envelope fluctuations in speech intelligibility, listeners have been provided with envelope frequencies well beyond a few tens of hertz in many speech studies (Eisenberg *et al.*, 2000; Qin and Oxenham, 2003; Apoux and Bacon, 2004; Gonzales and Oliver, 2005). One motivation for this is related to the relative uncertainty about precisely which envelope frequencies are most important for speech. Another motivation is that some speech information (e.g., periodicity cues) has been associated with envelope fluctuations between 50 and 500 Hz (Rosen, 1992). In studies simulating hearing via a cochlear implant (CI), listeners are commonly provided fast envelope fluctuations

(Dorman *et al.*, 1997, 1998; Friesen *et al.*, 2001; Faulkner *et al.*, 2003; Fu and Galvin, 2003; Baskent and Shannon, 2006). In this case, high-frequency cutoffs of the envelope filter (i.e., the low-pass filter used to extract the envelope) are used to conform to the cutoff frequencies commonly used in CI speech processors.

A. Factors affecting the bandwidth of the modulation spectrum

While it might seem reasonable to preserve high-frequency envelope fluctuations, it is unclear whether listeners can extract information from them. Even though humans are able to detect rates of amplitude modulation (AM) imposed on a broadband carrier as high as about 1000 Hz (Viemeister, 1979), filtering in the peripheral auditory system may limit the ability of the listeners to use high-frequency envelope cues in any given spectral region, as the maximum modulation frequency in a band-limited signal cannot exceed half the bandwidth of that band (see Lawson and Uhlenbeck, 1950). Modulation frequencies greater than half the bandwidth will be attenuated by the filter. Therefore, high-frequency AM should be available primarily from the high-frequency region of the speech spectrum. Similar limitations are also true for processing mimicking the resonant properties of the peripheral auditory system such as the analysis filters in CI and CI simulations (i.e., the filters used to decompose input signal spectra into a number of directly adjacent frequency bands for further processing). For example, Xu *et al.* (2005) examined the relative contribution of spectral and temporal cues for phoneme recognition by systematically varying the number of bands (1–16) and the cutoff frequency of the envelope low-pass filter (1–512 Hz) of a

^{a)}Author to whom correspondence should be addressed. Electronic mail: apoux@psycho.univ-paris5.fr

FIG. 1. Logarithm of the amplitude components of the modulation spectrum of a Gaussian noise, as computed at the output of a bandpass filter centered at 1000 Hz. The bandwidths of the bandpass filter were 1024 and 128 Hz for the upper and the lower panels, respectively. Prior to computation, the envelope of each band of noise was low pass filtered at 128 Hz.

noise-excited vocoder. According to their Table I, the bandwidth of the analysis filters ranged from 62 to 936 Hz in the 16-band condition. Therefore, it is clear that most, if not all, of the 16 unfiltered narrow-band envelopes could not exhibit fluctuations as high as 512 Hz. Moreover, if sidebands were introduced outside the passband of the analysis filter by AM, they would be significantly reduced by filtering after modulation (Eddins, 1993, 1999; Strickland and Viemeister, 1997). The effect of the bandwidth of the analysis filter is illustrated in Fig. 1. The two panels show the modulation spectrum of a Gaussian noise derived from the output of a bandpass filter centered at 1000 Hz. The bandwidth of the filter was 1024 Hz (upper panel) or 128 Hz (lower panel). The cutoff frequency of the envelope low-pass filter was 128 Hz in each condition. In the upper panel, it can be seen that only modulation frequencies above the cutoff of the envelope filter were attenuated. In contrast, modulation frequencies well below 128 Hz were attenuated when the bandwidth of the analysis filter was 128 Hz (lower panel). Only the modulation frequencies below half the bandwidth of the analysis filter (i.e., 64 Hz) were left intact.

B. Differential contribution of temporal envelope cues across frequency

From the above, it is apparent that the range of available envelope frequencies should differ across frequency because of the characteristics of the auditory filters and/or the analysis filters. Moreover, it is well established that the shape of the modulation spectrum of speech differs significantly

across frequency. Greenberg *et al.* (1998) indicated that the high-frequency spectral region of speech contains a significantly greater amount of energy in the midfrequency modulation spectrum (10–25 Hz) than does the lower frequency spectral region. More recently, Crouzet and Ainsworth (2001) showed that temporal envelopes extracted from distant frequency regions are only partially correlated. Consistent with previous reports (Plomp, 1983; Houtgast and Steeneken, 1985), the results of Crouzet and Ainsworth indicate that for very low-frequency envelope modulations (≤ 4 Hz) the correlation of envelope information between frequency bands is very high (>0.8). This high correlation reflects the syllabic rate in speech (i.e., syllables onsets and offsets are similar across the speech spectrum). For higher frequency envelope modulations (>8 Hz), the correlation remains significant only for adjacent spectral bands. Taken together, the fact that the shape of the modulation spectrum is not similar across spectral frequency and that narrow-band envelopes are only partially correlated across all spectral regions strongly suggest that temporal cues extracted from remote regions do not convey the same information.

Several behavioral studies with normal-hearing listeners reported data that are in good agreement with the above physical analysis. Grant *et al.* (1991) evaluated the benefits of the auditory presentation of envelope cues as a supplement to lip reading. Among the parameters assessed in this study, the authors investigated the effect of the frequency of the envelope low-pass filter and the effect of the center frequency of the analysis band. When the analysis band was located in the low-frequency region (500 Hz, 1 octave wide), some benefits to lip reading were observed for envelope cut-off frequencies up to 50 Hz. Benefits were observed up to 200 Hz when the analysis band was located in the high-frequency region (3150 Hz, 1 octave wide), suggesting a differential contribution of envelope cues across frequency. This result, however, may simply reflect the influence of auditory filters on the bandwidth of the modulation spectrum. More recently, Silipo *et al.* (1999) measured the intelligibility of one to four $\frac{1}{3}$ -octave-wide bands of speech (sentences) presented simultaneously. The envelope of one or two of the bands was low pass filtered in a systematic fashion, allowing the authors to assess the importance of low-, mid-, and high-frequency modulations at different spectral regions. Their results indicated that (i) only modulation rates below 12 Hz play a significant role in the low spectral region and (ii) modulation rates between 10 and 25 Hz are of particular importance for encoding speech information in the spectral region above 1.5 kHz. Finally, Apoux and Bacon (2004) assessed the relative importance of temporal information in broad spectral regions for consonant identification. For the purpose of forcing listeners to use primarily temporal envelope cues, speech sounds were spectrally degraded using four noise-band vocoder processing. Frequency-weighting functions were determined using two methods. The first consisted of measuring the intelligibility of speech with a hole in the spectrum either in quiet or in noise. The second consisted of correlating performance with the randomly and independently varied signal-to-noise ratio within each band. Results demonstrated that all bands contributed equally to consonant

identification when presented in quiet. In noise, however, both methods indicated that listeners consistently placed relatively more weight on the highest frequency band.

The first experiment of the present study was designed to demonstrate the effect of auditory filter bandwidth on the transmission of high-frequency AM. As discussed above, modulation frequencies greater than half the critical bandwidth (see Zwicker *et al.*, 1957) are presumably attenuated. Therefore, it is inappropriate to discuss the effects of envelope cutoff frequency when the effective modulation bandwidth is, in fact, limited by the bandwidth of the auditory filters. To evaluate the effect of the auditory filter bandwidth on the transmission of high-frequency AM, a vocoder-like processor with the cutoff frequency of the envelope low-pass filter in each band proportional to the critical bandwidth of the auditory filter corresponding to that band was implemented. More specifically, envelope cutoffs were independently determined for each band so that, in theory, the desired modulation frequencies in a particular band could not be attenuated by the auditory filters (or the analysis filters). The results were compared with those obtained using a single high-frequency envelope cutoff for all bands. For completeness, low-frequency envelope cutoff frequencies were also tested.

The second experiment was designed to independently establish for different frequency regions which AM rates are critical for speech intelligibility. Previous studies evaluating the importance of low- and high-frequency amplitude fluctuations for speech intelligibility always filtered the temporal envelope in each speech band homogeneously across frequency (i.e., the same envelope cutoff frequency was used for each band). In one study, however, the authors investigated the critical AM for speech in three broad frequency regions independently (Christiansen and Greenberg, 2007). The three bands were centered at 750, 1500, and 3000 Hz and were presented either in isolation or combined with one or two other bands. The envelope cutoff frequency in each band was 3, 6, 12, or 24 Hz and was always the same across bands when two or more bands were simultaneously presented. From the data reported in their Table I, it seems that the effect of the envelope filter was similar across bands when presented in isolation. The authors, however, did not perform a statistical analysis of their percent correct data, making it difficult to determine whether or not the same range of AM was critical in each spectral region. In the second experiment of the present study, the envelope cutoff frequency in one band was systematically varied while keeping the envelope cutoff frequency for all the other bands fixed. Because it was anticipated that performance may not be affected by the manipulation of the envelope cutoff frequency in a single band, the cutoff frequency in two adjacent bands was also systematically and uniformly varied while keeping the envelope cutoff frequency for all the other bands fixed.

A potential complicating factor in the interpretation of the results may arise from recent studies suggesting that envelope filtering techniques might be inappropriate to estimate the range of the pertinent fluctuation rates in speech. Indeed, a series of theoretical and behavioral studies indicated that envelope filtering techniques might introduce artifacts when

the speech fine structure is preserved (Ghitza, 2001; Atlas *et al.*, 2004). The exact reasons why envelope filtering techniques might be inappropriate are beyond the scope of the present study and, therefore, the reader is referred to a recent paper by Atlas *et al.* (2004) for further details. In short, it has been demonstrated that part of the original envelope information remains in the speech fine structure and that listeners are able to extract and use this information (Zeng *et al.*, 2004; Gilbert and Lorenzi, 2006). Accordingly, the possibility exists that listeners were, in fact, presented with rich envelope information in studies where the speech fine structure was preserved. Therefore, the speech fine structure was replaced by sinusoids in the present experiments. The use of spectrally degraded stimuli should also force listeners to primarily use temporal cues and, therefore, listeners likely should be more sensitive to modifications applied to the temporal structure of speech.

II. EXPERIMENT 1

A. Methods

1. Subjects

Six normal-hearing subjects participated—five females and one male [one of the authors, Apoux (S1), participated]. Their ages ranged from 20 to 32 years (average=23 years). Normal hearing was defined as having pure-tone air-conduction thresholds of 20 dB HL or better (ANSI, 1996) for octave frequencies from 250 to 8000 Hz. All participants, who, except for Apoux, were native American English speakers, had about 2 h of previous experience identifying spectrally degraded speech sounds and were paid an hourly wage for their participation.

2. Speech processing

The frequency range of speech stimuli was initially restricted to 0.1–5 kHz. The fine structure was replaced by sinusoidal carriers in the following manner. Each vowel-consonant-vowel (VCV) was first band passed in to N semi-logarithmic¹ frequency bands using 24th-order Butterworth filters ($N=4, 8, 12$, or 16). At the output of each analysis filter, the envelope was extracted by half-wave rectification and low-pass filtering at cf_m (eighth-order Butterworth). In all conditions but one, the envelope low-pass filter cutoff frequency was identical for each band ($cf_m=4, 8, 16, 160$, or 400 Hz). In the last condition, cf_m was independently computed for each band so that it was equal to half the critical bandwidth (HCB) of the auditory filter centered at the sinusoidal carrier frequency (Glasberg and Moore, 1990). The center frequency of each band and the corresponding envelope filter cutoff frequency in the HCB condition are given in Table I. The filtered envelopes were then used to modulate sinusoids with frequencies equal to the center frequencies of the bands. Each modulated sinusoid was passed through the same bandpass filter used in the original analysis band and the outputs of each band were finally combined. The overall level of each stimulus was normalized and calibrated to produce an average A-weighted output level of 70 dB.

TABLE I. Center frequencies of the analysis bands (in Hz). The envelope cutoff frequency—half the critical bandwidth—used for each band in the HCB condition is shown in *italics*.

Number of bands	Band															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
4	183	487	1 294	3 440												
	22	38	82	198												
8	132	214	350	570	930	1 517	2 473	4 033								
	19	24	31	43	62	94	146	230								
12	119	165	229	317	439	609	843	1 168	1 619	2 243	3 107	4 305				
	19	21	24	29	36	45	58	75	99	133	180	244				
16	114	145	186	237	303	387	494	630	805	1 028	1 313	1 676	2 141	2 734	3 491	4 458
	18	20	22	25	28	33	39	46	56	68	83	103	128	160	201	253

3. Speech material and procedure

The stimuli consisted of 16 consonants in /a/-consonant-/a/ environment recorded by four speakers (two for each gender) for a total of 64 VCV disyllables (Kwon and Turner, 2001). Each run or block presented the entire set of VCV stimuli and corresponded to a given combination between a band condition (4, 8, 12, or 16 bands) and an envelope cutoff condition (4, 8, 16, 160, 400 Hz or HCB). Therefore, each listener completed 24 blocks in which the 64 VCV stimuli were presented in random order. All conditions were presented in random order to avoid order effects. Stimulus presentation was computer controlled. Percent correct identification was measured using a single-interval, 16-alternative forced-choice procedure. The listeners were instructed to report the perceived consonant and responded using the computer mouse to select 1 of 16 buttons on the computer screen. Prior to data collection, each listener received 2 h of practice. No feedback was given during the practice or the experimental sessions. The listeners were tested individually in a double-walled, sound-attenuated booth. Stimuli were played to the listeners binaurally through Sennheiser HD 250 Linear II circumaural headphones.

B. Results and discussion

Figure 2 shows the percent correct identification score as a function of the number of bands averaged across listeners for each of the six envelope cutoff frequency conditions. The standard deviation across listeners was calculated for each condition, but for clarity these are not displayed. The standard deviation ranged from 6 to 21 percentage points (mean=13 percentage points). It generally increased with both increasing envelope cutoff and number of bands. Overall, the percent correct identification scores were slightly lower than in previous experiments. This apparent discrepancy, however, can be accounted for by differences in terms of the amount of practice. For example, Shannon *et al.* (1995) obtained about 90% correct but their subjects received between 8 and 10 h of practice. In studies with a similar amount of practice (e.g., 3 h in Xu *et al.*, 2005), the results were much more comparable. Other differences such as speech material and number of speakers may also account for these small discrepancies. Consistent with previous re-

ports, consonant identification generally improved as a function of the number of bands. Intelligibility also improved as the envelope cutoff frequency increased. Beyond 160 Hz, however, performance seemed to reach an asymptote: scores in the 160- and 400-Hz conditions were within 5 percentage points of one another. Although it was systematically lower, performance in the HCB condition was very similar to that in the 160- and 400-Hz conditions. Overall, the envelope cutoff frequency exerted a more dominant effect on consonant identification than did the number of bands. A repeated measure analysis of variance (ANOVA) with factors of number of bands and envelope cutoff frequency revealed a significant effect for both factors ($p < 0.001$). The interaction between the number of bands and the envelope cutoff frequency was also significant [$F(15, 143) = 2.21$, $p < 0.05$]. *Post hoc* (Tukey) comparisons confirmed that increasing the envelope cutoff frequency from 4 to 8 Hz, from 8 to 16 Hz, and from 16 to 160 Hz significantly improved the performance of the listeners ($p < 0.001$). Performance in the 400-Hz and HCB conditions did not differ significantly ($p = 0.30$). The difference between the 4-band and the 8-band conditions was significant ($p < 0.02$). The difference between the 8-band and the 12-band conditions was also significant ($p = 0.04$), while

FIG. 2. Mean percent scores for consonant identification as a function of the number of bands. The parameter is the cutoff frequency of the envelope low-pass filter.

the difference between the 12-band and the 16-band conditions was not ($p=0.47$). This last result is fairly consistent with previous reports showing that asymptotic performance is usually achieved with eight bands for VCV stimuli presented in quiet (Dorman *et al.*, 1997).

Also consistent with previous research (van Tasell *et al.*, 1987; Shannon *et al.*, 1995), our results indicate that usable speech information is contained in fluctuation rates above 16 Hz. Indeed, providing modulation frequencies above 16 Hz markedly improved performance, irrespective of the number of bands that was available. Asymptotic performance was reached in the 160-Hz condition for all number of band conditions, suggesting that AM rates between 160 and 400 Hz do not play a significant role in speech intelligibility. Results in the HCB condition showed that consonant identification is not affected by eliminating AM rates greater than half the bandwidth of an auditory filter and, therefore, well below 160 Hz in some regions of the speech spectrum (see Table I).

Because performance did not improve when envelope fluctuations above HCB were provided to the listeners, it is reasonable to assume that, at least in quiet, the transmission of envelope fluctuations can be limited to those frequencies below half the bandwidth of the auditory filter in CI and CI simulations. The present results, however, do not provide much indication about the possible influence of those frequencies above half the bandwidth of the auditory filter when speech is presented against a background noise. Possibly, the intelligibility of speech presented in noise might even decrease when fluctuation rates above half the bandwidth of the auditory filter are provided. First, our results showed that fluctuation rates above half the bandwidth of the auditory filter do not convey any usable information in terms of intelligibility. Therefore, limiting the transmission of non-pertinent envelope fluctuations should help reduce the amount of noise in CI users *without affecting intelligibility*. Second, interference in the modulation domain has been demonstrated for speech in at least two studies (Kwon and Turner, 2001; Apoux and Bacon, 2008). Accordingly, fluctuation rates *above* half the bandwidth of the auditory filter pertaining to the noise background might interfere with fluctuation rates *below* half the bandwidth of the auditory filter pertaining to the target. Since the latter play a role in speech intelligibility, performance might be adversely affected by such interference. In view of the above, we suggest that the transmission of envelope fluctuations should be limited to those frequencies below half the bandwidth of the auditory filter in CI and CI simulations, particularly when background noise is present.

III. EXPERIMENT 2

A. Conditions

Twelve new subjects participated in the second experiment (11 females and 1 male). Their ages ranged from 20 to 40 years (average=25 years). All participants were native American English speakers and had pure-tone air-conduction thresholds of 20 dB HL or better for octave frequencies from 250 to 8000 Hz. Subjects were paid an hourly

wage for their participation. The stimuli and processing were the same as for experiment 1 with the following exceptions.

- (1) Speech stimuli were processed through an eight-band processor only. This condition was chosen (i) to allow a reasonable spectral resolution (i.e., a large number of bands) while limiting redundancy of information between bands and (ii) because results from experiment 1 indicated that consonant identification should be around 50% correct in this condition with the 16 Hz envelope cutoff and, therefore, base line scores should be sufficiently removed from ceiling and floor effects.
- (2) In one condition (SING), the envelope cutoff frequency was set to 16 Hz for all but one band and the envelope in that single (test) band was low pass filtered at 4 Hz, 8 Hz, or HCB to quantify the range of modulation frequencies relevant for consonant identification in that particular band.
- (3) In the other condition (CONT), the envelope cutoff frequency was set to 16 Hz for all but two contiguous bands (1 and 2, 2 and 3, 3 and 4, and so on) and the envelope in each of these two (test) bands was low pass filtered at 4 Hz, 8 Hz, or HCB.
- (4) A baseline condition in which all narrow-band envelopes were filtered at 16 Hz was added. For comparison, a condition in which all eight narrow-band envelopes were filtered at half the critical bandwidth was also tested (HCB8).

The 12 listeners were divided into two groups. Each group first completed a series of seven blocks for practice. The first block corresponded to an unprocessed condition. The following six blocks corresponded to the conditions for which all bands were filtered at 16 Hz. We chose to train subjects in this particular condition because it was the reference condition. Then, the two speech band conditions were administrated separately. Six listeners completed the SING condition first and six completed the CONT condition first. The presentation order for the test band conditions was chosen randomly to avoid order effects. In addition, each listener completed three blocks in the 16-Hz condition and three blocks in the HCB8 condition. These 6 blocks were randomly dispersed among the 90 experimental blocks with the constraint to have at least one block for each condition in each part of the experiment.

B. Results and discussion

Figures 3 and 4 show the results for the SING and the CONT conditions, respectively. Each figure shows the performance averaged across the 12 listeners as a function of the test band(s) for each envelope cutoff condition. For reference, results for the baseline (all band at 16 Hz) and the HCB8 conditions are also reported. Note that individual performance in these two conditions corresponds to the average of 192 instead of 64 trials. The standard deviation across listeners was calculated for each condition, but for clarity these are not displayed. Overall, the standard deviation for a given condition was typically about 10 percentage points and did not vary systematically across conditions. Consistent

FIG. 3. Mean percent scores for consonant identification as a function of the band tested (SING condition). The parameter is the cutoff frequency of the envelope low-pass filter used for the tested band. Dashed and solid lines show the mean percent scores when the envelope filter for all bands was set to the same cutoff frequency, either 16 Hz or half the critical bandwidth (HCB8), respectively.

with our earlier finding, performance in the baseline condition was about 40% correct and increased up to about 60% correct in the HCB8 condition. First, consider the results for the SING condition (Fig. 3). Overall, the effects of test band frequency and envelope cutoff frequency were rather small. The latter had no influence on performance in the lower frequency region (test bands 1 and 2) for the most part. When the envelope cutoff was manipulated in test bands 3–8, some differences were observed. As a general rule, performance was only slightly reduced when the cutoff frequency was set to 4 or 8 Hz, while no effect was observed in the HCB condition. A repeated measures ANOVA with factors of test band frequency and envelope cutoff frequency showed a significant effect of test band frequency [$F(7,154)=2.84$, p

FIG. 4. Mean percent scores for consonant identification as a function of the bands tested (CONT condition). The parameter is the cutoff frequency of the envelope low-pass filter used for the tested bands. Dashed and solid lines show the mean percent scores when the envelope filter for all bands was set to the same cutoff frequency, either 16 Hz or half the critical bandwidth (HCB8), respectively.

<0.020] and envelope cutoff frequency [$F(2,154)=14.11$, $p<0.001$]. The interaction was also significant [$F(14,154)=2.47$, $p<0.005$]. *Post hoc* (Tukey) comparisons confirmed that the 4- and 8-Hz conditions did not differ significantly ($p=0.129$). However, performance in these two conditions was significantly lower than in the HCB condition ($p<0.001$ and $p<0.01$, respectively). *Post hoc* comparisons also revealed that scores for test bands 1 and 2 both differed significantly from performance for test band 4, indicating an effect of the frequency region ($p<0.03$ and $p<0.02$, respectively).

By comparing Figs. 3 and 4, it is apparent that the same general pattern was observed in the CONT condition. However, the effects were noticeably amplified by the manipulation of two contiguous bands simultaneously. Again, no difference was observed between the three envelope cutoff conditions in the low-frequency region (test bands 1&2). In the other test band conditions, restricting envelope fluctuations to frequencies below 4 or 8 Hz typically deteriorated performance relative to the base line condition with the 4-Hz condition leading systematically to poorer scores. In this latter condition, percent correct may be reduced by as much as about 20 percentage points (test bands 3&4 and 4&5). In a few band conditions, consonant identification scores in the HCB condition increased relative to those in the baseline condition. However, the difference did not exceed 5 percentage points. A repeated measure ANOVA with factors of test band frequency and envelope cutoff frequency showed a significant effect of test band frequency [$F(6,132)=7.79$, $p<0.001$] and envelope cutoff [$F(2,132)=54.55$, $p<0.001$]. The interaction was also significant [$F(12,132)=5.06$, $p<0.001$]. *Post hoc* comparisons indicated that all three envelope cutoff conditions differed significantly from each other. They also indicated that consonant identification scores in the test band 4&5 condition differed from test band 1&2, 2&3, 6&7, and 7&8 conditions.

Increasing the cutoff of the envelope low-pass filter to half the critical band of an auditory filter in one band only (SING) did not lead to any substantial improvement in intelligibility. This finding may seem surprising at first given the 20 percentage points advantage observed when all eight narrow-band envelopes were filtered at half the critical bandwidth (compare solid and dashed lines in Fig. 3). One likely explanation is that envelope information in each band interacted in a way that was not simply additive. Such synergistic interactions have been reported previously for narrow bands of speech. For example, Warren *et al.* (1995) examined the intelligibility of $\frac{1}{20}$ -octave-wide bands of speech presented either in isolation or in pairs. The authors found that two widely separated narrow bands of speech have intelligibility scores well above what would be predicted from the simple addition of their intelligibility scores when presented separately. The difference between the observed and the predicted performance, the synergistic effect, was about 17 percentage points in their study. To assess whether synergistic interactions also occurred in the CONT condition, the difference relative to the baseline condition in each test band condition was computed for both band conditions. A new condition, Sum Sing, was created in which the sum of differences ob-

FIG. 5. The upper, middle, and lower panels show the percent correct relative to the base line condition (all bands at 16 Hz) in the 4 Hz, 8 Hz, and HCB conditions respectively, as a function of the tested bands. In each panel, the filled and unfilled circles show the results for the sum Sing and the CONT condition, respectively.

tained for two bands was calculated for each pair of contiguous bands. For example, in the 1&2 band condition, the difference relative to the baseline condition computed for bands 1 and 2 was added together. This new condition allowed estimating the effect of manipulating two bands simultaneously (i.e., the synergistic effect). Figure 5 presents the results this way for the Sum Sing and the CONT conditions. The upper, middle, and lower panels show the results for the 4-Hz, 8-Hz, and HCB conditions, respectively. Positive values indicate an improvement over the baseline condition (cutoff frequency in all bands of 16 Hz). Statistical analyses were performed separately for each envelope cutoff frequency using repeated measures ANOVA with factors of test band frequency and “synergy” (i.e., Sum Sing versus CONT). The results indicated a significant effect of test band frequency for the 4- and 8-Hz conditions (all $p < 0.001$), confirming a differential contribution of rate between 4 and 8 and between 8 and 16 Hz to consonant identification. In contrast, the effect of test band frequency was not significant for the HCB condition. All three analyses indicated no significant synergistic effect. In other words, the improvement observed with two contiguous bands (CONT) was not larger

FIG. 6. Logarithm of the amplitude components of the modulation spectrum computed for all 64 VCVs used in the experiments as a function of the analysis band. The same eight bandpass filters used in the HCB condition were used for the analysis (i.e., the envelope low-pass filter cutoff frequency was equal to half the bandwidth of an auditory filter). Levels of gray indicate a 6 dB amplitude range. For clarity, lines on the modulation frequency axis corresponding to an envelope cutoff condition are in bold (4, 8, and 16 Hz).

than the summed increase measured individually for each band (Sum Sing). The interaction between the two main factors was not significant in all three analyses. A likely explanation for the absence of synergistic interactions with two contiguous bands is that adjacent bands provide highly correlated information. Therefore, the 20 percentage points advantage observed when all eight narrow-band envelopes were filtered at half the critical bandwidth most likely reflects the synergistic interaction of distant bands. This assumption is consistent with the results of a recent study by Healy and Warren (2003). In this study, the authors measured the intelligibility of band pairs presented either contiguously or separated by 1, 2, 3, or 4 octaves. Their results showed an improvement in sentence intelligibility as band pairs were moved apart. Although a band separation of 3 and 4 octaves resulted in no improvement relative to the contiguous condition, a 20 percentage points improvement was observed for band separation of 1 and 2 octaves.

An interesting question that emerges is how the present results can be related to the amount of “temporal information” that was physically present in each band. The modulation spectrum associated with each band is illustrated in Fig. 6, as computed for all 64 VCV stimuli at the output of the analysis filters used in the HCB condition. For clarity, the modulation frequencies corresponding to envelope cutoff conditions used in experiment 2 are marked by a bold line (4, 8, and 16 Hz). As can be seen, the peak of the spectrum lies around 2 Hz in all bands and the energy decreases rapidly with increasing modulation frequency beyond this value. The frequency of the peak presumably reflects the average duration of the stimuli. At most modulation frequencies, the

amount of energy is greatest in band 5 ($cf=930$ Hz) and lowest in bands 1, 7, and 8. The four remaining bands have an intermediate amount of energy with bands 4 and 6 being in the upper end of this intermediate range and bands 2 and 3 in the lower end. This pattern is fairly consistent with the effects of the envelope filter's cutoff frequency on consonant identification in that behavioral data indicated a primary role of modulation frequencies between 4 and 16 Hz in the mid-frequency region of the speech spectrum and a negligible contribution of these rates in the lower and upper frequency regions. Minor discrepancies may have resulted from at least two factors. First, it should be noted that a seven-point drop in percent correct can result from the deterioration of intelligibility for one VCV only. Therefore, the average identification scores may have been influenced by the results for a particular VCV or subset of VCVs and this effect might not be apparent when averaged behavioral and physical data are compared. Second, as discussed in the Introduction, the correlation of envelope information between adjacent bands is very high and, therefore, one can reasonably assume a certain degree of redundancy across bands. Accordingly, the loss of envelope information in a restricted frequency region may have been compensated by the presence of similar information in an adjacent band, reducing the observed effect. Comparisons across SING and CONT conditions provide at least some indirect support for this interpretation in that a larger effect of envelope cutoff was observed when the envelope cutoff was manipulated in two adjacent bands.

Results from experiment 2 suggest that the accuracy of methods developed for the prediction of speech intelligibility such as the speech transmission index (STI) (Steeneken and Houtgast, 1980) might be increased by measuring independently for different frequency regions which AM rates are critical for speech intelligibility. The STI is based on a weighted sum of modulation transfer functions, each of which specifies how modulation depth in seven octave-wide frequency bands is affected by interference or signal processing. This reduction in modulation depth of a test signal (Houtgast and Steeneken, 1985) or speech signal (Payton and Braida, 1999; Steeneken and Houtgast, 1999) has been shown to be highly correlated with intelligibility. Because STI calculation is based on the transmission of AM, one important parameter is the range of modulation frequencies considered. Based on the modulation spectra of octave-band filtered speech, modulation frequencies of 0.63–12.5 Hz in $\frac{1}{3}$ -octave steps have been typically used. According to our data, however, the upper limit of the pertinent fluctuations rates for speech intelligibility varies across the speech spectrum. Although our results were obtained with VCVs, it is reasonable to assume that the same range of modulation frequencies should not be used for each of the seven-octave-wide bands. In some bands, only the modulation frequencies below a few hertz should be considered. Conversely, modulation frequencies well above 12.5 Hz should be considered in other bands.

IV. SUMMARY AND CONCLUSIONS

The goal of the present study was to determine which modulation frequencies contribute significantly to the identi-

fication of spectrally degraded speech in specific frequency regions. The results from experiment 1 suggest that envelope fluctuations above 16 Hz play a critical role in consonant identification. They also demonstrate that the preservation of modulation frequencies greater than half the bandwidth of an auditory filter has no significant effect on consonant identification in quiet. Taken together, these findings indicate that most of the temporal information necessary to preserve accurate consonant identification is conveyed by envelope fluctuations equal or lower than half the critical bandwidth of an auditory filter. Although not surprising, this result stands out against the use of a high envelope cutoff frequency as typically used in CI and CI simulations.

The second experiment indicates that the range of critical modulation frequencies for consonant identification depends on the region of the speech spectrum. Behavioral data showed a primary role of modulation frequencies between 4 and 16 Hz in the midfrequency region of the speech spectrum and a negligible contribution of these rates in the lower and upper frequency region of the speech spectrum. Accordingly, previous studies showing that the removal of modulation frequencies between 4 and 16 Hz significantly affects speech intelligibility (Drullman *et al.*, 1994; Arai *et al.*, 1999) only reflect the importance of these rates in a specific region of the speech spectrum. A comparison between the analysis of the speech stimuli used in the present experiment and the behavioral data obtained with these stimuli suggests that understanding spectrally reduced speech (eight-band vocoder) relies on most, if not all, the available envelope fluctuations in each band. In other words, envelope fluctuations present in narrow bands of speech with enough power cannot be attenuated without affecting intelligibility.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Institute of Deafness and Other Communication Disorders (NIDCD Grant No. DC01376). We thank Ken Grant and two anonymous reviewers for their helpful comments on a previous version of the manuscript.

¹The upper cutoff frequency f_n of the n th band was computed according to the equation

$$f_n = F_{l0} \times 10^{\log_{10}(F_{up}/F_{l0})(1/N)n},$$

where F_{l0} and F_{up} are, respectively, the lower and upper limits of the filterbank in hertz (i.e., 100 and 5 000 Hz) and N is the total number of bands.

- ANSI (1996). "Specifications for audiometers," ANSI Report No. S3.6-1996, American National Standards Institute, New York.
- Apoux, F., and Bacon, S. P. (2004). "Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise," J. Acoust. Soc. Am. **116**, 1671–1680.
- Apoux, F., and Bacon, S. P. (2008). "Selectivity of modulation interference for consonant identification in normal-hearing listeners," J. Acoust. Soc. Am. **123**, 1665–1672.
- Arai, T., and Greenberg, S. (1997). "The temporal properties of spoken Japanese are similar to those of English," Proceedings of the Eurospeech, Rhodes, Greece, pp. 1011–1014, September.
- Arai, T., Pavel, M., Hermansky, H., and Avendano, C. (1999). "Syllable intelligibility for temporally filtered LPC cepstral trajectories," J. Acoust. Soc. Am. **105**, 2783–2791.

- Atlas, L., Li, Q., and Thompson, J. (2004). "Homomorphic modulation spectra," Proceedings of the 29th International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, pp. 761–764, May.
- Baskent, D., and Shannon, R. V. (2006). "Frequency transposition around dead regions simulated with a noiseband vocoder," *J. Acoust. Soc. Am.* **119**, 1156–1163.
- Christiansen, T. U., and Greenberg, S. (2007). "Distinguishing spectral and temporal properties of speech using an information-theoretic approach," XVIth International Congress of Phonetic Sciences, Saarbrücken, Germany, December.
- Crouzet, O., and Ainsworth, W. A. (2001). "On the various influences of envelope information on the perception of speech in adverse conditions: An analysis of between-channel envelope correlation," Workshop on Consistent and Reliable Cues for Sound Analysis, Aalborg, Denmark, September.
- 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.
- Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (1998). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels," *J. Acoust. Soc. Am.* **104**, 3583–3585.
- Drullman, R., Festen, J. M., and Plomp, R. (1994). "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.* **95**, 1053–1064.
- Eddins, D. A. (1993). "Amplitude modulation detection of narrow-band noise: Effects of absolute bandwidth and frequency region," *J. Acoust. Soc. Am.* **93**, 470–479.
- Eddins, D. A. (1999). "Amplitude-modulation detection at low- and high-audio frequencies," *J. Acoust. Soc. Am.* **105**, 829–837.
- Eisenberg, L. S., Shannon, R. V., Martinez, A. S., Wygonski, J., and Boothroyd, A. (2000). "Speech recognition with reduced spectral cues as a function of age," *J. Acoust. Soc. Am.* **107**, 2704–2710.
- Faulkner, A., Rosen, S., and Stanton, D. (2003). "Simulations of tonotopically mapped speech processors for cochlear implant electrodes varying in insertion depth," *J. Acoust. Soc. Am.* **113**, 1073–1080.
- Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (2001). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," *J. Acoust. Soc. Am.* **110**, 1150–1163.
- Fu, Q. J., and Galvin, J. J. (2003). "The effects of short-term training for spectrally mismatched noise-band speech," *J. Acoust. Soc. Am.* **113**, 1065–1072.
- Gilbert, G., and Lorenzi, C. (2006). "The ability of listeners to use recovered envelope cues from speech fine structure," *J. Acoust. Soc. Am.* **119**, 2438–2444.
- Ghitza, O. (2001). "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *J. Acoust. Soc. Am.* **110**, 1628–1640.
- Glasberg, B. R., and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**, 103–138.
- Gonzales, 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., Braida, L. D., and Renn, R. J. (1991). "Single band amplitude envelope cues as an aid to speechreading," *Q. J. Exp. Psychol. A* **43**, 621–645.
- Greenberg, S., Arai, T., and Silipo, R. (1998). "Speech intelligibility derived from exceedingly sparse spectral information," International Conference on Spoken Language Processing, Sydney, Australia, pp. 74–77, December.
- Greenberg, S. (1999). "Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation," *Speech Commun.* **29**, 159–176.
- Healy, E. W., and Warren, R. M. (2003). "The role of contrasting temporal amplitude patterns in the perception of speech," *J. Acoust. Soc. Am.* **113**, 1676–1688.
- 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.
- Kwon, B. J., and Turner, C. W. (2001). "Consonant identification under maskers with sinusoidal modulation: Masking release or modulation interference?," *J. Acoust. Soc. Am.* **110**, 1130–1140.
- Lawson, J. L., and Uhlenbeck, G. E. (1950). *Threshold Signals*, Radiation Laboratory Series Vol. 24 (McGraw-Hill, New York).
- Payton, K. L., and Braida, L. D. (1999). "A method to determine the Speech Transmission Index from speech waveforms," *J. Acoust. Soc. Am.* **106**, 3637–3648.
- Plomp, R. (1983). "The role of modulation in hearing," in *Hearing—Physiological bases and psychophysics*, edited by R. Klinke and R. Hartmann (Springer-Verlag, New York).
- Qin, M. K., and Oxenham, A. J. (2003). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," *J. Acoust. Soc. Am.* **114**, 446–454.
- Rosen, S. (1992). "Temporal information in speech: Acoustic, Auditory and Linguistic aspects," *Philos. Trans. R. Soc. London, Ser. B* **336**, 367–373.
- Silipo, R., Greenberg, S., and Arai, T. (1999). "Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representation," Proceedings of the Eurospeech, Budapest, Hungary, pp. 2687–2690, September.
- Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Steeneken, H. J. M., and Houtgast, T. (1980). "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.* **67**, 318–326.
- Steeneken, H. J. M., and Houtgast, T. (1999). "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.* **67**, 318–326.
- Strickland, E. A., and Viemeister, N. F. (1997). "The effects of frequency region and bandwidth on the temporal modulation transfer function," *J. Acoust. Soc. Am.* **102**, 1799–1810.
- van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widen, G. P. (1987). "Speech waveform envelope cues for consonant recognition," *J. Acoust. Soc. Am.* **82**, 1152–1161.
- Viemeister, N. F. (1979). "Temporal modulation transfer functions based upon modulation thresholds," *J. Acoust. Soc. Am.* **66**, 1364–1380.
- Warren, R. M., Riener, K. R., Bashford, J. A., and Brubaker, B. S. (1995). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," *Percept. Psychophys.* **57**, 175–182.
- Xu, L., Thompson, C. S., and Pöfingst, B. E. (2005). "Relative contributions of spectral and temporal cues for phoneme recognition," *J. Acoust. Soc. Am.* **117**, 3255–3267.
- Zeng, F.-G., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y.-Y., and Chen, H. (2004). "On the dichotomy in auditory perception between temporal envelope and fine structure cues," *J. Acoust. Soc. Am.* **116**, 1351–1354.
- Zwicker, E., Flottorp, G., and Stevens, S. S. (1957). "Critical band width in loudness summation," *J. Acoust. Soc. Am.* **29**, 548–557.