The ability of listeners to use recovered envelope cues from speech fine structure

Gaëtan Gilbert[a] and Christian Lorenzi
Laboratoire de Psychologie de la Perception, CNRS, Ecole Normale Supérieure, DEC, 29 rue d'Ulm, France, and Groupement de Recherche en Audiologie Expérimentale et Clinique, France

(Received 27 June 2005; revised 16 January 2006; accepted 17 January 2006)

Recent work has demonstrated that auditory filters recover temporal-envelope cues from speech fine structure when the former were removed by filtering or distortion. This study extended this work by assessing the contribution of recovered envelope cues to consonant perception as a function of the analysis bandwidth, when vowel-consonant-vowel (VCV) stimuli were processed in order to keep their fine structure only. The envelopes of these stimuli were extracted at the output of a bank of auditory filters and applied to pure tones whose frequency corresponded to the original filters' center frequencies. The resulting stimuli were found to be intelligible when the envelope was extracted from a single, wide analysis band. However, intelligibility decreases from one to eight bands with no further decrease beyond this value, indicating that the recovered envelope cues did not play a major role in consonant perception when the analysis bandwidth was narrower than four times the bandwidth of a normal auditory filter (i.e., number of analysis bands $\geqslant 8$ for frequencies spanning 80 to 8020 Hz). © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2173522]

PACS number(s): 43.71.Gv, 43.71.Es, 43.66.Mk [KWG] Pages: 2438–2444

I. INTRODUCTION

A number of speech perception studies have investigated the role of two temporal features of the speech signal in speech understanding: its slow varying component, referred to as the signal's envelope, and its fast varying component, referred to as the signal's fine structure. Several signal-processing techniques allow the extraction of the signal's fine structure while removing the temporal envelope; among them, infinite compression and the Hilbert transform (which leads to a decomposition of the signal into its envelope and fine structure) are the most commonly used. Based on these techniques, studies conducted by Licklider and Pollack (1948), Drullman *et al.* (1994a, b), and Smith *et al.* (2002) have shown that the removal of the temporal envelope while keeping the fine structure intact does not affect strongly speech intelligibility, when the processing is applied within a single, wide frequency band (also called the "analysis band") or a limited number of broad frequency bands (i.e., six, 1-oct bands). However, when the speech fine structure is extracted within 24 $\frac{1}{4}$-oct frequency bands, Drullman *et al.* (1994a, b) and Drullman (1995) showed that the intelligibility is degraded substantially.

Modelling work by Ghitza (2001) has provided an insight into this apparent discrepancy by demonstrating that the degraded speech-envelope cues may be recovered at the output of auditory filters because the signal's envelope and instantaneous frequency information are related. The notion that the degraded envelope cues may be recovered at the output of auditory filters and used by listeners was recently confirmed by Zeng *et al.* (2004). They conducted a speech identification task where the contribution of the recovered envelope to speech identification was assessed. First, they removed the whole envelope of the sentences using the Hilbert decomposition in a single, wide frequency band (80–8020 Hz). Then, they extracted the "recovered" envelopes of the processed sentences at the output of a bank of gammachirp auditory filters (Irino and Patterson, 1997). They finally used these envelopes to amplitude modulate noise bands having the same bandwidth as the original auditory filters. In agreement with Ghitza (2001)'s predictions, the resulting processed sentences were found to be intelligible (40% mean correct identification).

Zeng *et al.* (2004) pointed out that the envelope recovery process depends upon the ratio between the analysis bandwidth (i.e., the bandwidth of the filters used to analyze the speech stimuli) and the bandwidth of auditory filters. The following simulation illustrates this dependency for elementary stimuli. The test signal was obtained by adding two 100% sinusoidally amplitude-modulated (SAM) tones having equal peak amplitude (65 dB SPL), A and B. A was a 1-kHz tone modulated at 4 Hz and B was a 2-kHz tone modulated at 8 Hz. The starting modulation phase of each SAM tone was chosen at random. The dotted line in Fig. 1 shows the envelope of the test signal at the output of a 1-ERB-wide auditory filter centered at 1 kHz. As expected, this envelope fluctuates at a 4-Hz rate. The solid line shows the envelope recovered from the fine structure of the test signal at the output of this auditory filter centered at 1 kHz when the fine structure of the test signal was extracted[1] from the compound stimulus (i.e., A+B). The dashed line shows the envelope recovered from the fine structure of the test signal at the output of the same auditory filter when the fine structure of the test signal was extracted from each SAM tone (i.e., A and B) taken separately, the two fine structure signals being added thereafter. The second case (referred to

[a]Present address: G. Gilbert. MRC Institute of Hearing Research, Glasgow Royal Infirmary, Queen Elizabeth Building, 16 Alexandra Parade, Glasgow G31 2ER, UK. Electronic mail: gaetan@ihr.gla.ac.uk

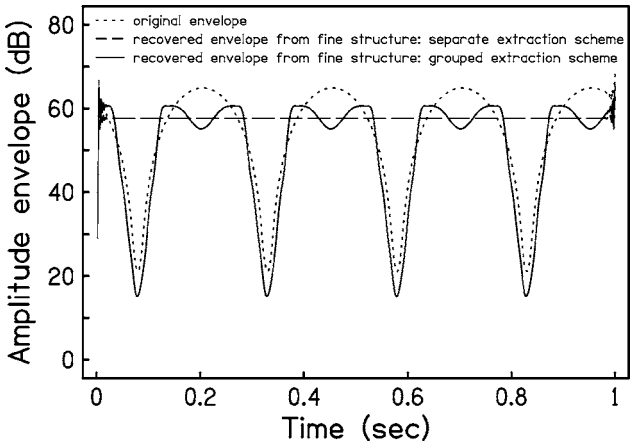

FIG. 1. Comparison between the envelopes obtained at the output of an auditory filter centered at 1 kHz (bandwidth of 1 ERB) for an elementary test signal. The test signal was obtained by adding two 100% sinusoidally amplitude-modulated (SAM) tones with equal peak amplitude (65 dB SPL): A and B. A was a 1-kHz tone modulated at 4 Hz and B was a 2-kHz tone modulated at 8 Hz. Dotted lines show the envelope of the test signal at the output of an auditory filter centered at 1 kHz. The dashed line shows the envelope recovered at the output of the same auditory filter from the fine structure of the test signal, when the fine structure of the test signal was extracted from each SAM tone taken separately, the two fine structure signals being added thereafter ("separated extraction" processing scheme). The solid line shows the envelope recovered at the output of the auditory filter centered at 1 kHz from the fine structure, when the fine structure was extracted from the compound stimulus ("grouped extraction" processing scheme).

as the "separate extraction scheme") corresponded to narrow analysis filtering while the first one (referred to as the "grouped extraction scheme") corresponded to broad analysis filtering. Figure 1 indicates clearly that envelope recovery occurs when the fine structure of the test signal was extracted from the compound stimulus. In other words, the broader the analysis bandwidth, the closer the recovered envelope is to the true envelope of the 1-kHz component. This suggests that when the bandwidth of analysis filters is substantially larger than the bandwidth of auditory filters, the envelope cues in those filters are mostly recovered.

In the current example, envelope recovery (solid line in Fig. 1) originated from an FM-to-AM conversion mechanism. As shown by Hartmann (1998), the beating of two (or more) frequency components creates a frequency modulation (FM) which depends on the relative amplitude of each component of the test signal (i.e., A and B): when the amplitude of tone A dominates over that of tone B, the instantaneous frequency of the test signal is closer to the frequency of tone A and vice versa. Thus, when the fine structure was extracted from the compound stimulus, that is when the analysis filter was broad, the variation over time of the amplitude ratio between the components A and B could be retraced from the FM. Such a FM was then converted into dynamic variations in the level of excitation (i.e., into an AM signal) at the output of auditory filters because the frequency excursion of the FM (within 1 and 2 kHz) was large compared to the bandwidth of auditory filters.

The goal of the present study was to extend the initial work by Zeng *et al.* (2004) by assessing the capacity of listeners to identify speech on the basis of the putative recovered envelope cues. This capacity was evaluated as a function of the analysis bandwidth, when envelope cues were removed by using Hilbert transform. A control identification task was conducted when the envelope was removed by infinite peak clipping.

II. EXPERIMENTS

A. Method

1. Speech material

One set of 48 unprocessed vowel-consonant-vowel (VCV) stimuli was recorded. These speech stimuli consisted of three exemplars of the 16 /aCa/ utterances (C $=/p,t,k,b,d,g,f,s,\int,m,n,r,l,v,z,ʒ/$) read by a French female speaker in quiet (mean VCV duration=648 ms; standard deviation=46 ms). Each signal was digitized via a 16-bit analog/digital converter at a 44.1-kHz sampling frequency.

2. Stimuli in the main identification task

The original speech signals were submitted to two different processing schemes. Stimuli processed using the first scheme (Hilbert fine structure conditions: HFS) contained speech information in their fine structure only, but envelope cues were potentially recoverable at the output of auditory filters. Stimuli were also generated using a second scheme (recovered envelope from Hilbert fine structure conditions: R-HFS) so as to force listeners to identify consonants primarily on the basis of the recovered envelope cues.

a. HFS conditions Each VCV signal was initially band-pass filtered using Butterworth filters (62 dB/oct rolloff) into 1, 2, 4, 8, or 16 complementary frequency bands (analysis bands) spanning the range 80–8020 Hz and following a logarithmic spacing inside this nominal bandwidth. Table I relates the number of frequency bands to their respective cutoff frequencies, and Table II relates the number of frequency bands to their bandwidths (in ERB units and in Hz). Forward and backward filtering were used to cancel phase delays. The Hilbert transform was then applied in each

TABLE I. Relationship between the number of frequency bands (i.e., analysis bands) and their respective cutoff frequencies.

No. of bands	Cutoff frequencies in Hz																
1	80																8020
2	80								801								8020
4	80				253				801				2535				8020
8	80		142		253		450		801		1425		2535		4509		8020
16	80	107	142	190	253	338	450	601	801	1068	1425	1900	2535	3380	4509	6013	8020

TABLE II. Bandwidth in ERB units as a function of the number of bands used in this study. The related bandwidth in Hertz is shown in italics below the value in ERB units.

No. of bands	Bandwidth of each band (in ERB units and *in Hz*) band number																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Average
1	30.5																30.5
	7940																
2	11.2	19.3															15.3
	721	*7219*															
4	4.1	7.1	9.2	10.1													7.6
	171	*548*	*1734*	*5845*													
8	1.7	2.1	3.2	3.9	4.4	4.8	5.0	5.1									3.8
	62	*111*	*197*	*351*	*624*	*1110*	*1974*	*3511*									
16	0.8	0.9	1.1	1.3	1.5	1.7	1.9	2.0	2.1	2.3	2.3	2.5	2.4	2.5	2.5	2.6	1.9
	27	*35*	*48*	*63*	*85*	*112*	*151*	*200*	*267*	*357*	*475*	*635*	*845*	*1129*	*1504*	*2007*	

band in order to decompose the VCV signal into its envelope (module of the Hilbert analytic signal) and temporal fine structure (cosine of the argument of the Hilbert analytic signal). The envelope component was discarded. The fine structure was multiplied by the root mean square (rms) power of the band-pass filtered VCV, in order to compensate for the reduction in amplitude caused by envelope removal. The "power-weighted" fine structure signals were finally summed over all frequency bands and presented as such to the listeners.

b. R-HFS conditions The HFS signals were passed through a bank of 30 gammachirp auditory filters, each 1 ERB wide (Irino and Patterson, 1997) with center frequencies ranging from 123 to 7743 Hz, and spaced along an ERB scale. In each band, the temporal envelopes were extracted using the Hilbert transform and low-pass filtered (cutoff frequency=64 Hz, 62 dB/oct rolloff) using a Butterworth filter (again forward and backward filtering were used). These envelopes were then used to amplitude modulate sine waves having the same frequencies as the original center frequencies of the auditory filters, but with random starting phase.

Figure 2 shows the correlation between the original and the recovered envelopes. The filled symbols show the mean correlation coefficients computed across the 48 VCV utterances between the speech envelopes of the original VCV stimuli and the envelopes of the HFS stimuli at the output of six auditory filters, as a function of the number of analysis bands.[2] The values were averaged across the 48 VCV utterances. A high correlation coefficient means that there was a close resemblance between the original envelope and that recovered at the output of auditory filters.

As predicted by Ghitza (2001), envelope cues were recovered from fine structure information at the output of auditory filters when the processing used to extract the fine structure was applied within a single, wide frequency band (filled circles). Figure 2 also shows that, overall, increasing the number of bands from one to eight (i.e., decreasing the bandwidth of each band) had a detrimental effect on envelope recovery. It is noteworthy that the envelope was better recovered for the auditory filter centered at 1197 Hz and that the detrimental effect of increasing the number of bands was especially large for this auditory filter. The upper panel of Fig. 3 shows that the average excitation pattern of the VCVs used in this study peaked in the frequency region around 1000 Hz (the frequency region around 1000 Hz encompassed the energy of the first and second formant of the vowel /a/ used in this study). The lower panel of Fig. 3 shows the amount of envelope recovery for the stimuli used in the single-band HFS condition (as in Fig. 2, but with a greater frequency resolution). It is quite clear from Fig. 3 that envelope recovery peaked in auditory filters where the amount of excitation is highest. Additionally, Fig. 2 shows slight negative correlation coefficients in auditory filters adjacent to the auditory filters conveying most of stimulus' energy. In auditory filters whose output is weak, the FM-to-AM conversion mechanism (responsible for envelope recovery) recovered mostly the true amplitude envelope from the auditory filters whose output is strongest; however, this recovered envelope is in antiphase with the true envelope. Knowing that the amplitude-envelopes of speech stimuli are, to some extent, positively correlated across the whole spec-

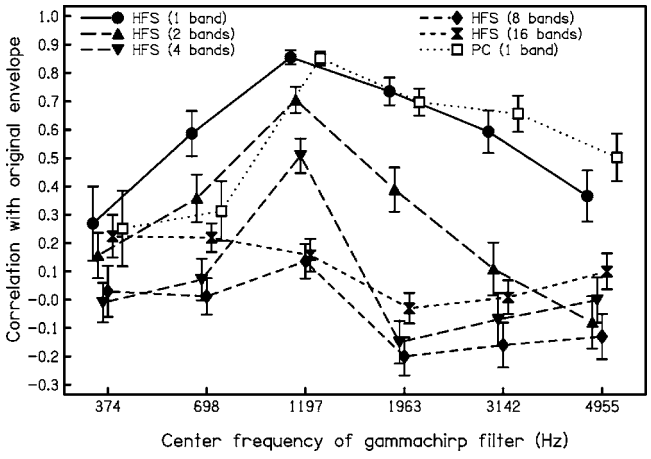

FIG. 2. Effect of the number of frequency bands (i.e., analysis bands) on the mean correlation coefficients computed (across the 48 VCV utterances) between the original speech envelopes and the envelopes of the stimuli in the Hilbert fine structure (HFS) conditions, at the output of six gammachirp auditory filters (filled symbols). Hilbert decomposition was used here to remove envelope cues while keeping fine structure intact. Open squares show the mean correlation coefficients computed between the original speech envelopes and the envelopes of stimuli in the peak clipping (PC) at the output of six gammachirp auditory filters, when infinite peak clipping in a single, wide frequency band is used to remove the temporal envelope. Error bars show ± one 95% confidence interval.

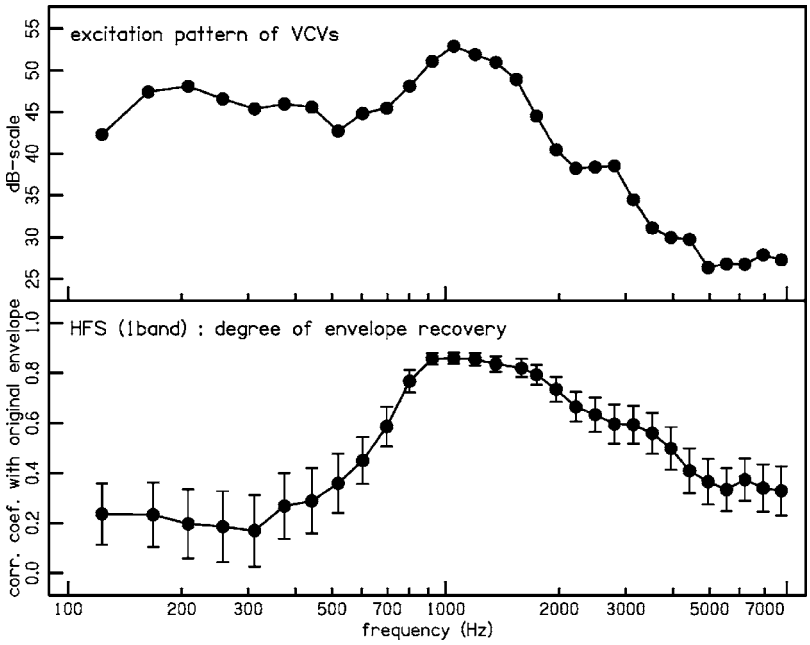

FIG. 3. The upper panel shows the excitation pattern averaged across the 48 VCV used in this study. Note the peak around 1 kHz, due to the first and second formant of the vowel /a/ used in this study. The lower panel shows the mean correlation coefficients computed between the original speech envelopes and the envelopes of the stimuli in the Hilbert fine structure (HFS) (1 band) condition at the output of 30 gammachirp auditory filters. The error bars in the lower panel represent ± one 95% confidence intervals. Note that the amount of envelope recovery is also peaking around 1 kHz.

trum (Crouzet and Ainsworth, 2001), this might explain the observation of those negative correlation coefficients.

3. Stimuli in the control identification task

An additional identification task was devised in order to assess to what extent the findings could be replicated when envelope cues were degraded using a different processing scheme. Indeed, in the case of *broadband* stimuli, the module and the argument of the analytic signal obtained by the Hilbert transform do not respectively correspond strictly to the envelope and instantaneous phase of the signal. Thus, fine structure information obtained by means of the Hilbert transform might be corrupted by the envelope information and vice versa (Ghitza, 2001). Said differently, the use of the Hilbert transform to demodulate broadband signals might be inadequate. In this additional identification task, infinite peak clipping in a single, wide frequency band was used to remove envelope cues from VCV signals (as in Licklider and Pollack, 1948). More precisely, each original signal was initially band-pass filtered (62 dB/oct rolloff, Butterworth filter, backward and forward filtering) between 80 and 8020 Hz. Infinite peak clipping was applied to each bandpass-filtered signal by replacing all positive amplitudes by +1 and all negative amplitudes by −1. Each clipped signal was then multiplied by the rms power of the VCV signal in that band (to compensate for the reduction in amplitude due to envelope removal). The resulting clipped stimuli were either presented as such (peak clipping condition: PC) or submitted to the second processing scheme (recovered envelope from peak clipping condition: R-PC), so as to force listeners to identify consonants primarily on the basis of the recovered envelope cues.

The open squares in Fig. 2 show the mean correlation coefficients computed across the 48 VCVs between the speech envelopes of the original VCVs and the envelopes of the infinite peak-clipped stimuli at the output of six auditory filters.[2] The correlation coefficients were similar to those obtained previously in the HFS single band condition, again indicating that envelope cues were recovered at the output of auditory filters. This reveals that the degree of envelope recovery reported in the broadband HFS conditions was not mainly due to an improper use of the Hilbert transform.

4. Procedure

All stimuli were generated using a 16-bit digital/analog converter operating at a sampling frequency of 44.1 kHz and delivered diotically via Sennheiser HD 580 earphones at an average level of 70 dB(A). The rms values of the stimuli were equalized. Listeners were tested individually in a sound-attenuating booth. In a typical experimental session, four complete and identical sets of the 48 VCV utterances corresponding to a given experimental condition (i.e., a given set of processed stimuli) were presented at random. Each listener was instructed to identify the presented consonant. The 16 possible choices were presented on the screen of the computer, and the listener entered his/her response by selecting a VCV on the screen with a computer mouse. No feedback was given to the listeners. The percentage of correct identification was calculated and a confusion matrix was built from the 192 VCV (4×48) utterances for a given set of stimuli.

All listeners were given at least 2 h of practice prior to data collection. In the main identification task, all sets of stimuli [2 experimental conditions (HFS vs. R-HFS) ×5 different number of bands (1 to 16 bands)] were presented in random order across listeners. In the control identification task, the two set of stimuli [two experimental conditions (PC vs. R-PC)] were presented in random order across listeners.

5. Listeners

A first group of five listeners participated in the main identification task. Their ages ranged from 20 to 36 years (mean age: 26 years; standard deviation: 6 years). A second group of five different listeners participated in the control

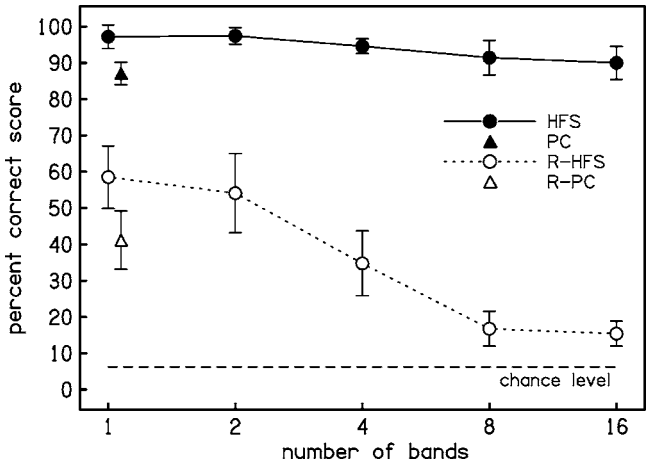

FIG. 4. Mean identification scores across listeners obtained in the Hilbert fine structure (HFS) (filled circles with continuous lines) and recovered envelopes from Hilbert fine structure (R-HFS) (open symbols with dotted lines) conditions. Hilbert decomposition was used here to remove envelope cues, while keeping fine structure intact. Data are plotted along with the mean data obtained in the peak clipping (PC) (filled triangle) and recovered envelopes from peak clipping (R-PC) (open triangle) conditions, where infinite peak clipping was used to remove envelope cues. Error bars show ± one 95% confidence interval. The dotted line shows chance level.

identification task. Their ages ranged from 20 to 22 years (mean age: 21 years; standard deviation: 1 year). All participants were native French speakers. They had audiometric pure tone thresholds less than 20 dB HL between 0.25 and 8 kHz and no history of hearing difficulty.

B. Results

1. Main identification task

Figure 4 shows the mean identification scores across listeners obtained in the HFS (filled circles with continuous lines) and R-HFS (open symbols with dotted lines) conditions. For all conditions, chance level corresponded to 6.25% correct.

The results showed that listeners identified perfectly or nearly perfectly consonants in the five HFS conditions (i.e., for 1 to 16 bands) with only a slight trend for performance to decrease (8%) as the number of bands increased from 2 to 16. The high scores observed in the present study contrast with the lower (70%–80%) scores obtained by Zeng *et al.* (2004) and Smith *et al.* (2002) using one-band, Hilbert fine structure sentences. This difference might be due to methodological differences between the three studies: in the present study, listeners received training before data collection and a closed-set format was used, whereas an open-set format was used in the two previous studies.

The identification scores measured in the R-HFS conditions were poorer than in the HFS conditions. They reached about 60% correct in the one- and two-band R-HFS conditions. This score was slightly higher than the 40% reported in a similar broadband condition by Zeng *et al.* (2004). This difference might be due to the same methodological differences listed above. This result indicates that the contribution of the recovered envelope cues to intelligibility in HFS conditions cannot be systematically neglected.

TABLE III. Phonetic features of the 16 French consonants used in this study (Martin, 1996).

Consonant	Voicing	Manner of articulation	Place of articulation
/p/	unoviced	occlusive	front
/t/	unvoiced	occlusive	middle
/k/	unvoiced	occlusive	back
/b/	voiced	occlusive	front
/d/	voiced	occlusive	middle
/g/	voiced	occlusive	back
/f/	unvoiced	constrictive	front
/s/	unvoiced	constrictive	middle
/ʃ/	unvoiced	constrictive	back
/v/	voiced	constrictive	front
/z/	voiced	constrictive	middle
/ʒ/	voiced	constrictive	back
/l/	voiced	constrictive	middle
/r/	voiced	constrictive	back
/m/	voiced	occlusive	middle
/n/	voiced	occlusive	front

However, Fig. 4 also shows that performance dropped significantly when the number of bands increased. A repeated measures analysis of variance (ANOVA) with factors number-of-bands and processing type confirmed this observation [main effect of factor number of bands: $F(4,16) = 32.12$, $p < 0.001$] [the percent correct identification scores were transformed into rationalized arcsine units (Studebaker, 1985) prior to the statistical analysis]. Nevertheless, this effect was much larger in the R-HFS condition than in the HFS condition [as shown by a significant interaction between factors processing condition and number of bands: $F(4,16) = 16.49$, $p < 0.001$]. Indeed, consonant identification in the R-HFS condition reached a minimum of about 15% in the 8–16 bands. Thus, a substantial contribution of recovered envelope cues to consonant identification was unlikely in the HFS condition when the stimuli were generated using 8 or 16 frequency bands.

Table II indicates that the bandwidth of these eight frequency bands varied from 1.7 ERB units for the lowest band to 5.1 ERB units for the highest band; the average bandwidth was 3.8 ERB units. This suggests that the ability to use recovered envelope cues for consonant identification was essentially abolished when the average analysis filters' bandwidth was less than approximately four times the bandwidth of a normal auditory filter.

The specific reception of the individual phonetic features of voicing (voiced versus unvoiced), manner (occlusive versus constrictive), and place (front versus middle versus back) was evaluated by an information-transmission analysis (Miller and Nicely, 1955) on the individual confusion matrices (see Table III for the assignment of the consonant features). The results of this analysis are presented in Fig. 5. The leftmost, middle, and rightmost panels show the results for the reception of voicing, manner, and place of articulation, respectively. In each panel, the filled circles with continuous lines and open circles with dotted lines correspond to the HFS and R-HFS conditions, respectively. The results show that the relatively good identification performance ob-

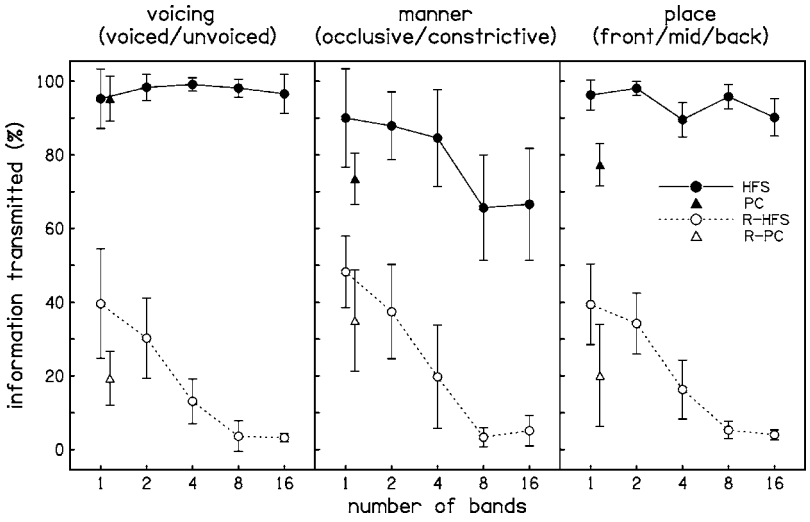

FIG. 5. Specific reception of three phonetic features as a function of the number of frequency bands. The rightmost, middle, and leftmost panels show the results obtained for the reception of voicing, manner, and place of articulation, respectively. In each panel, the filled circles with continuous lines and open circles with dotted lines correspond to the Hilbert fine structure (HFS) and recovered envelope from fine structure (R-HFS) conditions, respectively. Hilbert decomposition was used here to remove envelope cues, while keeping fine structure intact. Data are plotted along with the mean data obtained in the peak clipping (PC) (filled triangle) and recovered envelopes from peak clipping (R-PC) (open triangle) conditions where infinite peak clipping was used to remove envelope cues. Error bars show ± one 95% confidence interval.

served in the one- and two-band R-HFS conditions was due to a substantial reception of all three phonetic features (approximately 50% for the reception of manner, 40% for the reception of voicing and place in the single-band condition). Figure 5 shows that reception of voicing and place decreases when the number of bands increased from 1 to 16 in the R-HFS condition only. This is consistent with the observed significant interactions between the repeated factors number-of-bands and processing condition from two different repeated measures ANOVAs [in the case of voicing: $F(4,16)=28.91$; $p<0.001$; in the case of place: $F(4,16)=24.14$; $p<0.001$]. In contrast, reception of manner decreased when the number of bands increases from 1 to 16 in both experimental conditions (HFS and R-HFS), as also shown by the absence of interaction between the repeated factors number-of-bands and processing condition, from a repeated measures ANOVA performed on the transmitted information percentages in the case of manner [$F(4,16)=1.91$ $p=0.16$]. Thus, the slight decrease in identification performance with the frequency bandwith increasing observed in the HFS condition (cf. Fig. 4), was mainly caused by a decrease in the information transmitted by the manner of articulation. Moreover, the marked decrease in identification performance observed in the R-HFS condition (cf. Fig. 4) was caused by a decrease in the reception of all three phonetic features.

Taken together, the results presented in Figs. 4 and 5 demonstrate that the recovered envelope cues available at the output of auditory filters can contribute to consonant identification when the number of frequency bands used to extract the fine structure is four or less (i.e., when the average bandwidth of the analysis filters is greater than or equal to approximately 8 ERB units). However, when speech fine structure is extracted within eight bands or more (i.e., when the average bandwidth of the analysis filters is approximately smaller than or equal to 4 ERB units), recovered envelope cues are essentially abolished. These data also demonstrate that an *effective* removal of envelope and recovered envelope cues (i.e., when the number of bands is $\geqslant 8$ bands in HFS conditions) affects the transmission of manner, but not the transmission of voicing and place of articulation. This is consistent with previous work showing that nearly perfect reception of manner and poor reception of place are obtained when using mainly temporal envelope cues (e.g., Shannon *et al.*, 1995). However, this study also showed that nearly perfect reception of voicing can be obtained using mainly temporal envelope cues. This suggests that listeners can use either envelope or fine structure cues to reach nearly perfect reception of voicing; however, they must use essentially envelope cues to reach nearly perfect reception of manner and essentially fine structure cues to reach nearly perfect reception of place.

2. Control identification task

In the control identification task, infinite peak clipping was used to remove the amplitude envelope. This experiment was conducted in order to test whether or not the recovery of envelope cues in the broadband R-HFS conditions resulted from an inadequate use of the Hilbert transform to demodulate broadband signals (the Hilbert transform being, in theory, only applicable to demodulate signals that can be approximated as narrow band).

The mean identification scores across listeners obtained in the PC (filled triangle) and R-PC (open triangle) conditions are plotted in Fig. 4 along with the data of the first identification task. Overall, the data obtained with the single, wide band peak-clipped stimuli indicate that, as in Licklider and Pollack (1948), listeners identify nearly perfectly consonants (87% correct). Identification scores obtained in the R-PC condition (41% correct) were poorer than those measured in the PC condition. However, these scores were always substantially above chance level and comparable to those obtained in the main identification task using Hilbert decomposition. As shown in Fig. 5, the identification performance observed in the R-PC condition was, again, due to a substantial reception of all three phonetic features (approximately 35% for the reception of manner and 20% for the reception of voicing and place). This result confirms that the contribution of the recovered envelope cues to intelligibility, in conditions where speech items are processed within a single wide frequency band, cannot be neglected.

III. SUMMARY AND CONCLUSIONS

Taken together, the results indicate the following.

(1) Nearly perfect consonant identification can be obtained on the basis of speech fine structure cues extracted within 8 or 16 frequency bands. These data obtained with French material support the result of a recent study conducted by Xu and Pfingst (2003) with Mandarin material suggesting that speech identification in quiet does not rely entirely on temporal envelope cues.
(2) Consistent with Zeng *et al.*'s (2004) data, envelope cues recovered at the output of auditory filters can be used by listeners to identify consonants. When speech items are subjected to Hilbert decomposition or peak clipping within a single, wide frequency band so as to remove the amplitude envelope while keeping intact the speech fine structure, such a contribution is liable. This should be taken into account when interpreting the results obtained by Licklider and Pollack (1948), Drullman *et al.* (1994a, b), and Smith *et al.* (2002) in similar experimental conditions.
(3) The contribution of the recovered envelope cues to consonant identification is essentially abolished when the amplitude envelope of speech items is removed using Hilbert decomposition within analysis bands whose bandwidth is, on average, lower than or equal to four times the bandwidths of normal auditory filters, i.e., when the number of bands is equal to eight or more for frequencies ranging from 80 to 8020 Hz. This suggests that envelope recovery is unlikely to have significantly affected the results obtained by Drullman (1995), Drullman *et al.* (1994a, b), and Smith *et al.* (2002) when the fine structure of speech was extracted within eight frequency bands or more.

These results may help to process speech material in future studies attempting to compare speech identification based on envelope versus fine structure cues.

ACKNOWLEDGMENTS

This research was supported by a grant from ENTENDRE (GRAEC) to the first author, and a grant from the Institut Universitaire de France to C. Lorenzi. The authors wish to thank M. A. Akeroyd, K. W. Grant, F-G. Zeng, and one anonymous reviewer for valuable comments on preliminary versions of this manuscript. The authors are also thankful to I. Bergeras, D. Voillery, and C. Pinabiaux for their help in data collection.

[1]The fine structure of the tests signals corresponded to the cosines of the Hilbert analytic signals' arguments, multiplied by the original root mean square power of the test signals. Such a multiplication was required to correct for the reduction in amplitude due to envelope removal. The auditory filter used in the simulation was a gammachirp filter (Irino and Patterson, 1997).

[2]The envelopes were extracted using the Hilbert decomposition at the output of the gammachirp auditory filters. The resulting envelopes were passed forward and backward through a Butterworth low-pass filter (cutoff frequency=64 Hz, rolloff=62 dB/oct). A logarithmic transformation was applied before determining the correlation coefficient.

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.

Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592.

Drullman, R., Festen, J. M., and Plomp, R. (**1994a**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.

Drullman, R., Festen, J. M., and Plomp, R. (**1994b**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.

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.

Hartmann, W. (**1998**). *Signals, Sound, and Sensation* (Springer-Verlag, New York), Chap. 17, pp. 393–398.

Irino, T., and Patterson, R. D. (**1997**). "A time-domain level dependent auditory filter: The gammachirp," J. Acoust. Soc. Am. **101**, 412–419.

Licklider, J. C. R., and Pollack, I. (**1948**). "Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech," J. Acoust. Soc. Am. **20**, 42–51.

Martin, P. (**1996**). *Éléments de phonétique, avec application au français* (Les Presses de l'Université Laval, Sainte-Foy).

Miller, G. A., and Nicely, P. E. (**1955**). "Analysis of perceptual confusions among some english consonants," J. Acoust. Soc. Am. **27**, 338–352.

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

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature (London) **416**, 87–90.

Studebaker, G. A. (**1985**). "A rationalized arcsine transform," J. Speech Hear. Res. **28**, 455–462.

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.

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.