Role and relative contribution of temporal envelope and fine structure cues in sentence recognition by normal-hearing listeners

Frédéric Apoux,^{a)} Sarah E. Yoho, Carla L. Youngdahl, and Eric W. Healy

Speech Psychoacoustics Laboratory, Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210

(Received 2 January 2013; revised 27 June 2013; accepted 2 July 2013)

The present study investigated the role and relative contribution of envelope and temporal fine structure (TFS) to sentence recognition in noise. Target and masker stimuli were added at five different signal-to-noise ratios (SNRs) and filtered into 30 contiguous frequency bands. The envelope and TFS were extracted from each band by Hilbert decomposition. The final stimuli consisted of the envelope of the target/masker sound mixture at x dB SNR and the TFS of the *same* sound mixture at y dB SNR. A first experiment showed a very limited contribution of TFS cues, indicating that sentence recognition in noise relies almost exclusively on temporal envelope cues. A second experiment showed that replacing the carrier of a sound mixture with noise (vocoder processing) cannot be considered equivalent to disrupting the TFS of the target signal by adding a background noise. Accordingly, a re-evaluation of the vocoder approach as a model to further understand the role of TFS cues in noisy situations may be necessary. Overall, these data are consistent with the view that speech information is primarily extracted from the envelope while TFS cues are primarily used to detect glimpses of the target. © 2013 Acoustical Society of America.
[http://dx.doi.org/10.1121/1.4816413]

PACS number(s): 43.71.An, 43.71.Bp, 43.66.Ba, 43.66.Cb [EB]

Pages: 2205–2212

I. INTRODUCTION

Speech recognition in noise can be considerably disrupted by replacing the carrier or temporal fine structure (TFS) of the sound mixture with noise or tones in a series of frequency bands (i.e., vocoder processing). This disruption is typically larger if the background noise is fluctuating (e.g., Nelson *et al.*, 2003; Qin and Oxenham, 2003, 2006; Stickney *et al.*, 2005; Füllgrabe *et al.*, 2006; Gnansia *et al.*, 2009; Hopkins and Moore, 2009; Apoux and Healy, 2011). While evidence that replacing the original TFS negatively impacts the intelligibility of speech in noise is reasonably compelling, the implications of this vocoder processing may not be fully appreciated. The reduction in speech intelligibility associated with vocoder processing has typically been attributed to the elimination of acoustic speech cues potentially conveyed by the TFS. However, vocoder processing also results in the target and masker envelopes sharing an unrelated carrier. This carrier, typically a broadband noise or a tonal complex, would tend to be attributed to a single sound source by a listener if presented in isolation. It may therefore be argued that listeners are not being presented with a signal and a masker when there is only a single carrier. Moreover, it has been suggested that “TFS information could be important for identification of signals in masker dips, thus directing attention to the portions of a signal with the most favorable signal-to-noise ratio” (SNR; Hopkins and Moore, 2009). It is clear then how vocoder processing, by

eliminating all carrier disparities, might limit the listeners’ ability to detect the portions of a signal with the most favorable SNR. Because they tend to be attributed to a single sound source, the carriers used in vocoder processing will be referred to as single carriers throughout this paper. In contrast, carriers resulting from the mixture of two independent sounds, such as two voices, would tend to prompt the formation of two distinct auditory streams (e.g., Bregman and Campbell, 1971). These will be referred to as mixed carriers.

Another apparent limitation of vocoder processing is that the TFS of the sound mixture does not belong to either individual sound source from which the envelopes were extracted. Recent findings from studies investigating the envelope recovery phenomenon (Ghitza, 2001; Zeng *et al.*, 2004; Gilbert and Lorenzi, 2006; Apoux *et al.*, 2011) suggest that the use of an unrelated carrier may result in the introduction of considerable inconsistencies between the TFS and the envelope. In these studies, it was demonstrated that envelope information may be recovered from the TFS. In the case of unprocessed or natural speech stimuli, the envelope recovered from the TFS will match the “true” envelope (e.g., Heinz and Swaminathan, 2009). Intelligibility should therefore be unaffected or even reinforced in such conditions. In the case of vocoded speech stimuli, the recovered envelope will originate from the tonal- or noise-carrier TFS. Consequently, it will not match the true speech envelope. In these vocoded cases, strong inconsistencies will therefore exist between the envelope recovered from the TFS and the true envelope. These inconsistencies may provide conflicting cues and/or result in modulation masking and/or interference, therefore contributing to the difficulties experienced by normal-hearing (NH)

^{a)}Author to whom correspondence should be addressed. Electronic mail: fred.apoux@gmail.com

listeners when recognizing speech in noise under vocoded conditions. Taken together, the above limitations suggest that while vocoder processing may be a suitable approach when studying populations with no access to the original TFS, such as cochlear implant (CI) users, it may not be appropriate when studying NH or hearing-impaired (HI) listeners.

In the present study, the role and contribution of TFS cues to speech recognition in noise were investigated using an approach directly inspired by the signal processing described in [Smith *et al.* \(2002\)](#). In the [Smith *et al.*](#) study, the authors synthesized stimuli having the envelope of one sound and the TFS of another. They referred to these stimuli as chimeric sounds. In the present study, we synthesized stimuli having the envelope of one version of a sound mixture and the TFS of another version of the “same” sound mixture. In contrast to [Smith *et al.*](#), chimeric sounds were therefore synthesized using two versions of the same sound mixture. More specifically, the envelope and TFS were extracted from the same target and masker mixture, but the SNR at which these targets and maskers were mixed often differed. For instance, the target and masker may be mixed at 0 dB SNR to create the stimulus from which the envelope was obtained while the same target and masker were mixed at -12 dB SNR to create the stimulus from which the TFS was obtained. As a result, the SNR of the envelope and that of the TFS was 0 and -12 dB, respectively. The obvious advantage to this approach is that the TFS of the sound mixture is still composed of the two original carriers. A first consequence is that the original relationship between envelope and TFS in the chimeric sound is preserved, at least to some extent. A second consequence is that envelope recovery, the possibility to recover envelope information from the TFS, should not be a concern here because the original envelope is preserved to some extent. Moreover, the envelope, but more importantly the TFS, was extracted from 30 narrow frequency bands. As demonstrated by [Gilbert and Lorenzi \(2006\)](#), the use of narrow analysis filters to extract the TFS severely limits the possibility to recover envelope information from the TFS. Whereas the primary purpose of the present study was to investigate the contribution of TFS cues to speech recognition in noise, the contribution of envelope cues was also evaluated. More specifically, the present approach offered the opportunity to evaluate the relative contribution of envelope and TFS cues.

II. EXPERIMENT 1: INDEPENDENT MANIPULATION OF THE SNR

A. Methods

1. Subjects

Forty NH listeners participated in the first experiment (37 females). Their ages ranged from 19 to 26 yr (average = 21 yr). All participants had pure-tone air-conduction thresholds of 20 dB HL or better at octave frequencies from 250 to 8000 Hz ([ANSI, 2004, 2010](#)). They were paid an hourly wage or received course credit for their participation. This study was approved by The Ohio State University Institutional Review Board.

2. Speech material and processing

The target stimuli consisted of 350 sentences randomly selected from the Speech Perception In Noise (SPIN; [Kalikow *et al.*, 1977](#)) test. They included 175 low-predictability sentences and 175 high-predictability sentences from the standard male-voice recordings. The sentences were presented along with one of two masker types. One masker type consisted of a simplified speech spectrum-shaped noise (SSN, constant spectrum level below 800 Hz and 6 dB/octave roll-off above 800 Hz). The other masker type consisted of speech sentences randomly selected from the AzBio test (SPE; [Spahr *et al.*, 2012](#)).¹ The maskers were always gated on and off with the target sentence.

The stimuli were created as follows (see Fig. 1). Target and masker stimuli were added at two different SNRs: -12 to 12 dB and -18 to 6 dB in 6-dB steps for SSN and SPE, respectively. These values were selected according to a preliminary experiment to produce recognition scores that cover the range from floor to ceiling. After addition, target and masker stimuli were filtered into 30 contiguous frequency bands ranging from 80 to 7563 Hz using two cascaded 12th-order digital Butterworth filters (for details, see [Apoux and Healy, 2009](#)). Stimuli were filtered in both the forward and reverse directions so that the filtering process produced zero phase distortion. Each band was one equivalent rectangular bandwidth (ERB_N; [Glasberg and Moore, 1990](#)) so that the filtering roughly simulated the frequency selectivity of the normal auditory system. The envelope and TFS were extracted from each band using Hilbert decomposition as described in [Apoux *et al.* \(2011\)](#). The envelope extracted from a given band was then imposed on the TFS of the corresponding band obtained from the same stimulus mixed at a different SNR. The resulting amplitude-modulated TFS was then band-pass filtered using the initial analysis filter. The 30 amplitude-modulated TFSs were finally summed to create the final stimulus. It should be noted that the two sound mixtures from which the envelope and TFS were independently obtained differed only in their SNR (i.e., the same target and masker sentences were used). Thus the final stimuli consisted of the envelope of a sound mixture at x dB SNR and the TFS of the *same* sound mixture at y dB SNR. The SNR of the sound mixture from which the envelope was obtained will be referred to as SNR_{env}. The SNR of the sound mixture

FIG. 1. Schematic of the processing used to create the stimuli. The grayed area illustrates the processing within each of the 30 analysis filters.

from which the TFS was obtained will be referred to as SNR_{TFS} . The overall level of the 30 summed bands was normalized and calibrated to produce 65 dBA.

The effect of the above processing was examined by computing correlations between the original “clean” speech signal and the processed stimuli. This was done separately for envelope and TFS. The original speech signal corresponded to 50 concatenated sentences from the SPIN corpus. Correlations for each SNR from -18 to 18 dB in 3-dB steps were computed with the SSN and SPE maskers. Values within analysis filter bands were calculated separately then averaged. The results of this analysis are presented in Fig. 2. Not surprisingly, TFS correlations were generally lower than envelope correlations. This was particularly true at the most favorable SNRs. According to an additional analysis, perfect TFS correlation (>0.99) was only achieved around 60 and 54 dB SNR for SSN and SPE, respectively, whereas perfect envelope correlations only required 24 dB SNR, irrespective of the masker type. Correlations decreased gradually with decreasing SNR, reaching extremely low values (<0.2) at the lowest SNR. Despite these differences between envelope and TFS, a considerable and systematic change in correlation was observed for both cues over the range of SNRs tested, suggesting that envelope and TFS were both substantially affected by the current manipulations.

3. Procedure

The 40 participants were divided randomly into two groups. Each group was tested in one masker-type condition: SSN or SPE. None of the participants had prior exposure to the SPIN test. All combinations of SNR_{env} and SNR_{TFS} were tested, resulting in 24 conditions for each masker type. Participants were tested individually in a double-walled, sound-attenuated booth. Stimuli were presented diotically to the participants through Sennheiser HD 280 Pro circumaural headphones. The experiments were controlled using custom MATLAB routines running on PCs equipped with high-quality digital-to-analog converters (Echo Gina3G). Participants

were instructed to type the final word of each target sentence using the computer keyboard. A short practice was provided prior to data collection consisting of three blocks with each block corresponding to recognition of eight sentences not used during formal testing. During practice, the SNR was the same for the envelope and TFS and was set to 1000, 6, and -6 dB in the first, second, and third blocks, respectively. The practice masker type was the same as that used for data collection. Feedback was provided during practice but not during the experimental session. After practice, each participant completed the 24 experimental blocks in random order. Each experimental block corresponded to recognition of 14 sentences: Seven sentences randomly selected from the low-predictability list and seven sentences randomly selected from the high-predictability list. There was no replacement so that each sentence was heard only once. The total duration of testing, including practice, was approximately 2 h.

B. Results

The group mean scores for the SSN and SPE conditions are presented in Figs. 3 and 4, respectively. In each figure, the left panel shows the percentage of final words correctly reported as a function of SNR_{env} , while the right panel shows the same data as a function of SNR_{TFS} . The parameter is SNR_{TFS} in the left panel and SNR_{env} in the right panel. In each panel, a bold line connects the five data points obtained in the conditions in which the SNR_{env} and SNR_{TFS} were identical and therefore, corresponds to baseline performance on speech in noise (i.e., reference function; REF). As one may reasonably expect, baseline performance increased with increasing SNR, irrespective of the background type. It was also poorer in the SSN than in the SPE condition at comparable SNRs. The standard deviation across participants ranged from 0% to 15% points (mean = 11.9% points) in the SSN conditions and from 6% to 19% points (mean = 11.5% points) in the SPE conditions. For clarity, these are not displayed. As plotted, an effect of the factor represented on the abscissa would result in the functions having a strictly positive (or negative) slope. If this factor (SNR_{env} in the left panels and SNR_{TFS} in the right panels) was to account for most of the effect (i.e., the factor represented as parameter in the figure was irrelevant), the shape of the functions would be similar to that of the REF function and the functions would largely overlap. Although a generally similar pattern was observed in the two masker-type conditions, a few quantitative differences were noted. The results for each masker type are therefore reported separately.

Consider first the results obtained in the presence of the SSN masker (Fig. 3). As can be seen in the left panel of Fig. 3, the SNR of both the envelope and the TFS affected performance. The five functions show a positive slope, and their overall shape is generally similar to that of the REF function, indicating an effect of SNR_{env} . The slopes, however, are shallower at high SNR_{TFS} , suggesting a smaller influence of SNR_{env} . There was a substantial effect of SNR_{TFS} at low SNR_{env} (-12 and -6 dB) when the SNR of the TFS was strictly positive. In the remaining conditions, the effect of SNR_{TFS} was limited as indicated by the corresponding

FIG. 2. Correlation coefficients between the original speech and the chimeric sound's envelopes (filled symbols) and TFS (open symbols) as a function of the SNR. The circles correspond to the SSN conditions while the squares correspond to the SPE conditions.

FIG. 3. Average sentence recognition scores in speech-shaped noise (SSN) as a function of the SNR of the envelope (left panel) and as a function of the SNR of the TFS (right panel) with SNR_{tfs} and SNR_{env} as parameter, respectively. In each panel, a bold line (REF) connects the data points for which SNR_{env} and SNR_{tfs} were equal.

functions in the right panel being mostly horizontal. Overall, the effect of noise on sentence recognition could be attributed almost entirely to SNR_{env} in the -12 , -6 , and 0 dB- SNR_{tfs} conditions. Consider now the results obtained in the presence of the SPE masker (Fig. 4). In the left panel of Fig. 4, it can be seen that all five functions show a positive slope, indicating an effect of SNR_{env} in all conditions. More importantly, the shape of these functions matches almost exactly that of the reference function. The functions also overlap considerably, suggesting a limited effect of SNR_{tfs} . This limited effect of SNR_{tfs} is also apparent in the right panel of Fig. 4, as all five functions are essentially horizontal. Taken together, these results suggest that performance depended almost exclusively on the SNR of the envelope in the presence of the speech (i.e., fluctuating) background.

A repeated-measures analysis of variance (ANOVA) on rationalized arcsine units (RAU; Studebaker, 1985) with factors SNR_{env} and SNR_{tfs} was performed separately for each masker-type condition. These analyses confirmed a main effect of SNR_{env} in SSN [$F(4,16) = 701.3$, $p < 0.001$] and in SPE [$F(4,16) = 340.4$, $p < 0.001$]. The effect of SNR_{tfs} was also significant with both SSN [$F(4,16) = 146.7$, $p < 0.001$] and SPE [$F(4,16) = 7.4$, $p < 0.001$]. The two-way interactions were all significant ($p < 0.001$), suggesting a differential contribution of one structure (envelope or TFS) as a function of the SNR of the other structure. An analysis of effect size (Cohen, 1988)² was performed to further assess these effects. The magnitude of the SNR_{env} effect averaged 67% points across the five SNR_{tfs} and can be characterized as large ($d = 1.40$). Although the effect of SNR_{tfs} was less than 17% points on average, its magnitude could still be characterized as moderate ($d = 0.40$). A closer examination

of the data, however, revealed that two of the ten SNR_{tfs} functions (-12 and -6 dB SNR_{env} in SSN) yielded a large effect ($d = 1.21$). The remaining eight SNR_{tfs} functions yielded an effect that can be characterized as small ($d = 0.20$). Despite the statistically significant effect of SNR_{tfs} , it may therefore be concluded from Exp. 1 that distortion of the TFS had a limited effect on speech recognition, especially in the fluctuating background.

III. EXPERIMENT 2: NATURE OF THE TFS

Experiment 1 showed that, in most conditions, disturbing the TFS of the target signal does not have a large effect on sentence recognition. This finding contrasts with the large role of TFS reported in vocoder studies, especially when a fluctuating background is present. As mentioned previously, there are at least two important differences between the chimera and the vocoder approaches, both of which are related to the nature of the TFS. One difference is that following vocoder processing, the target and masker envelopes share a single carrier. While numerous demonstrations exist to show that the normal auditory system is capable of making inferences about missing portions of a target signal (e.g., Warren, 1970; Warren *et al.*, 1972; Warren *et al.*, 1997), it is unclear how it handles a missing carrier. However, it is reasonable to assume that imposing “multiple” envelopes on a single carrier may not improve intelligibility as it could affect the ability to extract the envelope of the target from that of the background. A second difference between the chimera and the vocoder approaches is related to the consistency between the TFS and the envelope(s). In our implementation of the chimera approach, envelope and TFS belong to the same

FIG. 4. The same as Fig. 3 but for the speech masker (SPE).

sound mixture while they are obtained from two independent signals in the vocoder approach. The lack of consistency between envelope and TFS resulting from vocoder processing may also contribute to poor intelligibility. For instance, the recovered envelope may interfere with the processing of the true envelope as suggested in Sec. I. These two hypotheses were further investigated in a second experiment. More specifically, the effect of having only one carrier was assessed by preserving only the TFS of the target or that of the masker. The consistency hypothesis was then assessed by comparing the effect of preserving only the TFS of the target or that of the masker to that of replacing the TFS with an independent noise.

A. Methods

1. Subjects

Twelve NH listeners participated in the second experiment (10 females) and were compensated as in Exp. 1. Their ages ranged from 20 to 22 yr (average = 20.8 yr). All participants had pure-tone air-conduction thresholds of 20 dB HL or better at octave frequencies from 250 to 8000 Hz (ANSI, 2004, 2010).

2. Stimuli and procedure

As in Exp. 1, the target sentences were selected from the SPIN test and presented with one of the two maskers. This second experiment consisted of two parts. All the subjects participated in the first part in which they were presented with stimuli created using the procedure described previously (Fig. 1). However, fewer SNRs were tested. The SNR_{env} was set to -6 or 0 dB for SSN and -18 , -12 , or 0 dB for SPE. In both masker-type conditions, SNR_{TFS} was set to -1000 (masker TFS only), 0 , or 1000 dB (target TFS only). We chose to use large SNR values instead of preserving only one carrier to keep the processing somewhat similar across conditions. Each combination of SNR_{env} and SNR_{TFS} was tested, resulting in six conditions for SSN and nine conditions for SPE.

Five of the 12 subjects also participated in the second part of the experiment, which consisted of four additional conditions. These additional conditions were randomly distributed among the 15 initial conditions for a total of 19

conditions. SNR_{TFS} was not manipulated in this part of the experiment, as the original TFS in each band was replaced with a speech-shaped noise (constant spectrum level below 800 Hz and 6 dB/octave roll-off above 800 Hz) created independently from the masker. In two of these additional conditions, SNR_{env} was set to -6 or 0 dB and the masker was SSN. In the other two additional conditions, SNR_{env} was set to -12 or 0 dB and the masker was SPE. Other methodological and procedural details were identical to those used in the first experiment.

B. Results

The data from the first part of Exp. 2 are presented in Fig. 5 (filled symbols). The left and right panels show the percentage of final words correctly reported as a function of SNR_{env} for the SSN and SPE maskers, respectively. The parameter is SNR_{TFS} . Selected data from the first experiment are replotted for comparison (open symbol). First, it should be noted that performance in the 0 -dB- SNR_{TFS} condition was nearly identical in the first and second experiments. The results obtained in the SSN condition (left panel) generally confirmed the findings of Exp. 1: A limited influence of SNR_{TFS} . The only effect, a modest drop in intelligibility, was observed in the -1000 -dB- SNR_{TFS} condition at -6 dB SNR_{env} . This is consistent with the influence of SNR_{TFS} observed at negative SNR_{env} in Exp. 1. The results obtained in the SPE condition (right panel) are more remarkable. Having SNR_{TFS} set to either -1000 or 1000 dB led to a generally comparable drop in performance at negative SNR_{env} .³ In other words, having only one carrier, whether it be the target or the masker, resulted in poor intelligibility in the presence of the single-talker background.

Figure 6 shows the noise carrier (i.e., vocoder) data obtained from five subjects who participated in both parts of Exp. 2 (black and white symbol). It also shows selected data from the first part of Exp. 2 obtained from the same five subjects (filled symbols).⁴ It is apparent from Fig. 6 that replacing the TFS with noise was generally more detrimental to sentence intelligibility than any of the other manipulations. The drop in performance resulting from vocoder processing was approximately 20% points in three out of the four SNR_{env} conditions when compared to preserving only the masker TFS (-1000 dB SNR_{TFS}). In the fourth condition

FIG. 5. Average sentence recognition scores as a function of the SNR of the envelope (SNR_{env}). The parameter is the SNR of the TFS (SNR_{TFS}). The left and right panels show scores in speech-shaped noise (SSN) and speech (SPE), respectively. In each panel, the filled symbols correspond to the data from Exp. 2, while the open symbol corresponds to selected data from Exp. 1.

FIG. 6. Average sentence recognition scores as a function of the SNR of the envelope (SNR_{env}). The left and right panels show scores in speech-shaped noise (SSN) and speech (SPE), respectively. In each panel, the filled symbols correspond to the data for three different SNR_{TFS} values while the black and white symbol corresponds to the noise-carrier data (i.e., vocoder).

(SSN; -6 dB SNR_{env}), performance was nearly identical. However, performance reached the floor in this last condition, limiting the possibility for recognition scores to drop further in the vocoder condition.

IV. GENERAL DISCUSSION

The main goal of the present study was to assess the relative contribution of envelope and TFS cues when both cues are available simultaneously⁵ in conditions that are more realistic than those obtained with the vocoder approach. The technique used to distort independently the envelope and the TFS also allowed circumvention of a few limitations of the vocoder approach used in previous studies. These limitations included the use of a single carrier to carry a mixture of envelopes and a lack of consistency between this single carrier and the mixed envelopes. Experiment 1 showed that distortion of the TFS alone does not substantially affect speech recognition in the presence of envelope information. In contrast, a strong correlation between distortion of the envelope and intelligibility was observed, suggesting that the identification of speech sounds relies almost exclusively on temporal envelope cues.

A similar chimera approach was recently used by Fogerty (2011). In this study, the author investigated the perceptual weight of envelope and TFS cues using the correlational method (e.g., Doherty and Turner, 1996; Apoux and Bacon, 2004). It was found that the SNR of the envelope and that of the TFS both contributed to sentence recognition. The relative weights for envelope and TFS were equal in one experiment and formed roughly a 2:1 ratio in another. This stands in some contrast to the results of the current experiments in which intelligibility was driven more strongly by envelope. However, some of this apparent discrepancy may be attributed to the fact that the analysis bands used by Fogerty (2011) were substantially wider than those employed here and therefore did not match the frequency selectivity of the normal auditory system. In the previous study, the stimuli were divided into three broad bands

spanning 80–6400 Hz. This difference in bandwidth is critical as it has been demonstrated that envelope recovery is especially pronounced when wide analysis bands are used (Gilbert and Lorenzi, 2006). Therefore, it is possible that the relative weights obtained for TFS were inflated in Fogerty (2011) due to the recovery of envelope from the TFS. In such cases of envelope recovery, performance can be driven by the recovered envelope under conditions in which TFS is presumed to be the dominant cue. This interpretation is in accord with that proposed by Ghitza (2001) and also accounts for results such as those reported in Smith *et al.* (2002), showing a substantial contribution of the TFS to sentence recognition with wide analysis bands.

While the role of TFS was found to be significant in the present study, it is apparent that it was very limited, especially in the presence of the competing talker. This limited role of TFS in fluctuating backgrounds contrasts appreciably with the results of previous work as the present study showed a larger effect of disrupting the TFS in the presence of a steady background, whereas previous studies typically reported a larger effect in the presence of a fluctuating background. Experiment 2 demonstrated that this discrepancy could be attributed to the use of a single carrier to carry a mixture of envelopes and the lack of consistency between this single carrier and the mixed envelopes in the vocoder approach as initially hypothesized.

First, Exp. 2 showed that having only one carrier is particularly detrimental to sentence intelligibility. While the reduced performance obtained at the low SNR (i.e., -1000 dB SNR_{TFS}) could reasonably be attributed to the presence of only the masker TFS, such interpretation is not possible for the 1000 -dB- SNR_{TFS} condition. Indeed, this latter condition was equally detrimental to sentence intelligibility despite preserving only the target TFS. Taken together, these results demonstrate that having only one carrier, *even if it is that of the target*, is more detrimental to speech recognition than preserving the mixed TFS (i.e., both target and masker TFSs). These findings have implications for our understanding of how the normal auditory system uses envelope and TFS cues. The fact that preserving only the target-speech TFS is more detrimental to speech recognition than preserving the mixed TFS suggests that NH listeners do not extract speech information from the TFS.⁶ Indeed, if the target-speech TFS was to provide speech information, then better performance should be observed when it is presented in isolation. This suggestion is also supported by the fact that preserving only the target-speech TFS is equivalent to preserving only the masker TFS. Instead, this latter finding suggests that the auditory system relies on TFS cues to segregate the target from the masker. More specifically, it is suggested that NH listeners use TFS cues to determine which auditory filter outputs are dominated by the target signal. When the original (i.e., mixed) TFS is replaced by that of the target or that of the masker, TFS-based segregation is no longer possible. This interpretation is consistent with the suggestion of Hopkins and Moore (2009) mentioned earlier. It is also consistent with the larger role of TFS in fluctuating backgrounds in that fluctuating backgrounds provide more and better opportunities to access the target TFS than do

steady backgrounds. It is not surprising then that performance drops more substantially in fluctuating backgrounds when TFS cues are removed. Finally, the use of a single carrier may be considered more generally “misleading,” as the auditory system may interpret such stimuli as originating from a single sound source.

Second, Exp. 2 showed that the consistency, or lack thereof, between the envelope and the TFS of the sound mixture can also be detrimental to sentence intelligibility. In the present study, the envelope/TFS relationship was varied systematically by manipulating SNR_{env} and SNR_{TFS} independently. In some conditions, both SNRs were similar and the envelope/TFS consistency was optimum. In other conditions, the SNRs differed by as much as 1000 dB and envelope/TFS consistency was rather poor. Finally, the TFS was replaced with noise in a last condition and envelope/TFS consistency did not exist. The lack of consistency may not have been as critical in the SSN condition as in the SPE condition. Data from Exp. 2, especially from the SPE masker, showed a gradual effect of envelope/TFS consistency in that sentence intelligibility decreased as consistency between envelope and TFS decreased. In view of the present results, a re-evaluation of the vocoder approach as a model to further our understanding of the role of TFS cues in noisy situations may be necessary. In particular, it is apparent from Exp. 2 that replacing the carrier of a sound mixture with noise cannot be considered equivalent to disrupting the TFS of the target signal.

The results of the present study suggest a dichotomy between the role of temporal envelope and TFS cues in speech recognition. Consistent with the speech transmission index (STI; Steeneken and Houtgast, 1980) and the results of many vocoder studies, it is suggested that the identification of speech sounds relies almost exclusively on temporal envelope cues. In general accord with the results of previous TFS studies, our data showed a statistically significant but limited role of TFS cues. These results may be interpreted in the context of the glimpsing model of speech recognition in noise (Cooke, 2006; see also Apoux and Healy, 2009, 2010 for a more specific discussion). In this model, speech recognition in noise relies on the ability to extract speech information (i.e., acoustic speech cues) from time-frequency regions in which the target signal is relatively unaffected by the background, the so-called glimpses. It is still unclear, however, which cues are used to detect or select these time-frequency regions. By showing an adverse effect of eliminating the disparities between target and masker(s) TFS, the present findings suggest that the auditory system may use TFS cues for segregation purposes. Accordingly, it is proposed here that the primary role of TFS is to assist in identifying which auditory channels are dominated by the target signal so that the output of these channels can be combined at a later stage to reconstruct the internal representation of that target (i.e., across-channel integration).

While NH listeners did not seem to extract speech information from the TFS in the present experiments, studies by Lorenzi and colleagues have shown otherwise. They reported almost perfect consonant identification by NH listeners when presented with the TFS only (Gilbert and

Lorenzi, 2006; Lorenzi *et al.*, 2006). These results may be interpreted as evidence that TFS does convey speech information. However, an alternate explanation is that TFS-based speech recognition is mainly supported by envelope recovery. This explanation is consistent with the results of Apoux *et al.* (2011) showing that Hilbert decomposition does not remove envelope information from the TFS and that NH listeners may recover this information at the output of the auditory filters. In any case, it may be assumed that for real-world listening situations where both the envelope and TFS are available, speech information is essentially extracted from the envelope. This suggestion is consistent with two particular results of Smith *et al.* (2002). First, these authors showed that NH listeners can perfectly identify the sentence coded by the envelope when listening to speech-speech (envelope-TFS) chimeras under processing conditions that best mimic normal frequency selectivity. Conversely, they showed that listeners have essentially no access to the sentence coded by the TFS in those same conditions. Second, Smith *et al.* (2002) reported that listeners can identify the sentence coded by the envelope more accurately in speech-noise (envelope-TFS) chimeras than in speech-speech chimeras. One may argue that this poorer identification when the TFS is a distracter sentence than when it is noise is the result of the additional modulation masking and/or interference produced by the non-target speech TFS, by way of envelope recovery. Again this suggests that when the original envelope is present, recognition is primarily based on the true envelope (as opposed to recovered envelope).

The present data have the potential to further our understanding of the difficulties experienced by CI patients when recognizing speech in noise. Based on the aforementioned TFS studies, it has been suggested that CIs should provide TFS information. The current data are in some accord with this suggestion in that they show a benefit of preserving the original TFS. The current data, however, suggest an alternate motive for preserving the original TFS. Indeed, we indicated earlier that most speech information is extracted from the envelope and that TFS might be primarily used as a grouping cue to determine which auditory filters convey undistorted target envelope cues. Accordingly, preserving the *original* TFS may not be critical for speech recognition in noise. Instead, CIs should process the incoming sound mixture so as to generate a unique carrier for each constituent source. It is apparent how not having to preserve the original TFS would simplify processing. Moreover, CI speech processors are already capable of segregating a target sound from the background and of generating various simple carriers. Another computational benefit is that this approach limits the need to determine which source is the target and which source is the masker as the goal is no longer to eliminate noise. It may be predicted that the addition of TFS cues will improve the auditory system’s ability to segregate sources. In other words, this approach should allow a “natural” and hopefully better extraction of the target from the background as it will provide some of the missing cues for glimpsing. Concomitantly, because no sound source is suppressed, the proposed technique would allow CI users to switch between sources as each source would be preserved.

ACKNOWLEDGMENTS

This research was supported by grants from the National Institute on Deafness and Other Communication Disorders (NIDCD Grant No. DC009892 awarded to author F.A. and DC008594 awarded to author E.W.H.).

¹Two of the five talkers were used. They were chosen so that their average fundamental frequency did not overlap with that of the target speaker (120 Hz). One talker, a male, was chosen based on a lower fundamental (93 Hz). The other talker, a female, was chosen based on a higher fundamental (190 Hz).

²Cohen's d was calculated based on the mean of the effect sizes at incremental SNRs using RAU-transformed scores and pooled standard deviations.

³The possibility that the smaller drop in performance observed at -18 dB SNR_{env} was due to a floor effect cannot be excluded.

⁴A comparison between Figs. 5 and 6 suggests that the pattern of results for these five listeners was slightly different from that of the entire group. To assess the extent of this apparent difference, we compared the mean data from these five listeners with that of the group (i.e., 12 listeners). On average, the five listeners were within 3.5 percentage points of the group means.

⁵It is assumed that the availability of temporal envelope and fine structure cues is dependent on the SNR.

⁶It should be noted that this suggestion is primarily valid for conditions where envelope cues are also available. This, however, should encompass most real-world situations.

ANSI (2004). S3.21 (R2009), *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (Acoustical Society of America, New York).

ANSI (2010). S3.6-2010, *American National Standard Specification for Audiometers* (Acoustical Society of America, 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 Healy, E. W. (2009). "On the number of auditory filter outputs needed to understand speech: Further evidence for auditory channel independence," *Hear. Res.* **255**, 99–108.

Apoux, F., and Healy, E. W. (2010). "Relative contribution of off- and on-frequency spectral components of background noise to the masking of unprocessed and vocoded speech," *J. Acoust. Soc. Am.* **128**, 2075–2084.

Apoux, F., and Healy, E. W. (2011). "Relative contribution of target and masker temporal fine structure to the unmasking of consonants in noise," *J. Acoust. Soc. Am.* **130**, 4044–4052.

Apoux, F., Millman, R. E., Viemeister, N. F., Brown, C. A., and Bacon, S. P. (2011). "On the mechanisms involved in the recovery of envelope information from temporal fine structure," *J. Acoust. Soc. Am.* **130**, 273–282.

Bregman, A. S., and Campbell, J. (1971). "Primary auditory stream segregation and the perception of order in rapid sequences of tones," *J. Exp. Psychol.* **89**, 244–249.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Earlbaum, Hillsdale, NJ), pp. 23–27.

Cooke, M. P. (2006). "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.* **119**, 1562–1573.

Doherty, K. A., and Turner, C. W. (1996). "Use of a correlational method to estimate a listener's weighting function for speech," *J. Acoust. Soc. Am.* **100**, 3769–3773.

Fogerty, D. (2011). "Perceptual weighting of individual and concurrent cues for sentence intelligibility: Frequency, envelope, and fine structure," *J. Acoust. Soc. Am.* **129**, 977–988.

Füllgrabe, C., Berthommier, F., and Lorenzi, C. (2006). "Masking release for consonant features in temporally fluctuating background noise," *Hear. Res.* **211**, 74–84.

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.

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.

Glasberg, B. R., and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**, 103–138.

Gnansia, D., Péan, V., Meyer, B., and Lorenzi, C. (2009). "Effects of spectral smearing and temporal fine structure degradation on speech masking release," *J. Acoust. Soc. Am.* **125**, 4023–4033.

Heinz, M. G., and Swaminathan, J. (2009). "Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech," *J. Assoc. Res. Otolaryngol.* **10**, 407–423.

Hopkins, K., and Moore, B. C. J. (2009). "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," *J. Acoust. Soc. Am.* **125**, 442–446.

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (1977). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *J. Acoust. Soc. Am.* **61**, 1337–1351.

Lorenzi, C., Gilbert, G., Cam, H., Garnier, S., and Moore, B. C. J. (2006). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proc. Natl. Acad. Sci. U.S.A.* **103**, 18866–18869.

Nelson, P. B., Jin, S.-H., Carney, A. E., and Nelson, D. A. (2003). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," *J. Acoust. Soc. Am.* **113**, 961–968.

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.

Qin, M. K., and Oxenham, A. J. (2006). "Effects of introducing unprocessed low-frequency information on the reception of envelope-vocoder processed speech," *J. Acoust. Soc. Am.* **119**, 2417–2426.

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

Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., Loisel, L. M., Oakes, T., and Cook, S. (2012). "Development and validation of the AzBio sentence lists," *Ear Hear.* **33**, 112–117.

Steeneken, H. J. M., and Houtgast, T. (1980). "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.* **67**, 318–326.

Stickney, G. S., Nie, K., and Zeng, F.-G. (2005). "Contribution of frequency modulation to speech recognition in noise," *J. Acoust. Soc. Am.* **118**, 2412–2420.

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

Warren, R. M. (1970). "Perceptual restoration of missing speech sounds," *Science* **167**, 392–393.

Warren, R. M., Hainsworth, K. R., Brubaker, B. S., Bashford, J. R., and Healy, E. W. (1997). "Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps," *Percept. Psychophys.* **59**, 275–283.

Warren, R. M., Obusek, C. J., and Ackroff, J. M. (1972). "Auditory induction: Perceptual synthesis of absent sounds," *Science* **176**, 1149–1151.

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.