Relative contribution of target and masker temporal fine structure to the unmasking of consonants in noise

Frédéric Apoux^{a)} and Eric W. Healy

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

(Received 11 January 2011; revised 2 August 2011; accepted 25 September 2011)

The present study assessed the relative contribution of the “target” and “masker” temporal fine structure (TFS) when identifying consonants. Accordingly, the TFS of the target and that of the masker were manipulated simultaneously or independently. A 30 band vocoder was used to replace the original TFS of the stimuli with tones. Four masker types were used. They included a speech-shaped noise, a speech-shaped noise modulated by a speech envelope, a sentence, or a sentence played backward. When the TFS of the target and that of the masker were disrupted simultaneously, consonant recognition dropped significantly compared to the unprocessed condition for all masker types, except the speech-shaped noise. Disruption of only the target TFS led to a significant drop in performance with all masker types. In contrast, disruption of only the masker TFS had no effect on recognition. Overall, the present data are consistent with previous work showing that TFS information plays a significant role in speech recognition in noise, especially when the noise fluctuates over time. However, the present study indicates that listeners rely primarily on TFS information in the target and that the nature of the masker TFS has a very limited influence on the outcome of the unmasking process. © 2011 Acoustical Society of America. [DOI: 10.1121/1.3652888]

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

Pages: 4044–4052

I. INTRODUCTION

In recent years, many studies have investigated the role of temporal fine structure (TFS) in speech recognition by comparing the intelligibility of natural—or unprocessed—speech with that of speech processed to replace the original TFS with tones or noise bands (i.e., vocoder processing; Shannon *et al.*, 1995). Results from these studies typically showed better speech recognition in noise for stimuli whose original TFS had been preserved. Although a benefit from preserved TFS is usually observed with most masker types, the nature of the masker, and more specifically the nature of the masker’s envelope, may considerably influence the extent of this benefit. Indeed, the drop in intelligibility associated with vocoder processing is generally larger for maskers whose envelope fluctuates over time, suggesting a greater role of TFS when extracting speech from temporally fluctuating backgrounds. For example, Hopkins and Moore (2009) measured speech reception thresholds for unprocessed and vocoded sentences in the presence of a steady or amplitude-modulated (AM) speech-shaped noise. While a benefit from preserved TFS was observed with both noise types, the authors found a larger benefit in the presence of the AM noise (6 dB) than in the presence of the steady noise (2 dB). The apparent interaction between vocoder processing and fluctuations in the masker observed in this study and others (e.g., Nelson *et al.*, 2003; Qin and Oxenham, 2003, 2006; Stickney *et al.*, 2005; Fullgrabe *et al.*, 2006; Gnansia *et al.*, 2009) has led to the suggestion that the auditory system uses TFS to identify the portions of a sound mixture containing relevant information about the target signal. In

other words, TFS information would support speech recognition in noise by “revealing” which portions of a sound mixture have a more favorable signal-to-noise ratio (SNR).

The apparent interaction between vocoder processing and envelope fluctuations in the masker may also be interpreted as evidence that TFS provides useful information to identify speech sounds. The presence of dips in the background would then allow listeners to have better access to this information. Such an interpretation is supported by the results of a set of studies in which speech recognition in noise for normal-hearing (NH) listeners was measured using stimuli processed solely to remove the temporal envelope (Gilbert and Lorenzi, 2006; Lorenzi *et al.*, 2006; Sheft *et al.*, 2008).¹ In other words, subjects were presented with stimuli that principally contained TFS information. Results showed that NH listeners can achieve high levels of speech understanding when presented with the TFS of speech stimuli, suggesting that TFS conveys sufficient information to accurately identify speech sounds. Because the extent to which envelope information can be recovered from the TFS is still a matter of debate, the results of these studies should be interpreted with caution (Ghitza, 2001; Apoux *et al.*, 2011).

A recent study by Apoux and Healy (2010a) attempted to circumvent the above limitation by removing the original TFS information using a more conventional approach, a vocoder-like approach. However, one important limitation of this approach is that performance would remain at ceiling if the TFS information only was removed from the stimuli. To avoid this potential ceiling effect, the authors divided the speech spectrum into 30 contiguous one equivalent rectangular bandwidth (ERB_N; Glasberg and Moore, 1990) bands and presented only a subset of bands. In one condition, the TFS of the speech bands was left intact and in another condition, it was replaced with tones using vocoder processing.

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

Replacing the speech TFS with tones led to a systematic drop in performance, confirming the contribution of TFS cues to the identification of speech sounds.

According to the previous discussion, the auditory system may use TFS cues (a) to determine which portions of a sound mixture have a relatively favorable SNR or (b) to directly identify speech sounds. It should be noted that these two mechanisms are not mutually exclusive and could each potentially account for a portion of the TFS benefit. In any case, it is of interest to note these mechanisms may primarily rely on the TFS in the target. In the case of mechanism (b), it is apparent that only the TFS in the target provides information about the identity of the target speech. In the case of mechanism (a), there is no direct evidence supporting a primary role of the TFS in the target. However, the fact that the effect of preserved TFS is larger in fluctuating backgrounds suggests that TFS information is better used during the masker dips. In other words, it is when the representation of the masker TFS is poorest (i.e., in the masker dips) that the normal auditory system takes the most advantage of TFS information. Finally, [Gnansia et al. \(2008\)](#) showed that masking release for vocoded stimuli increases with increasing modulation depth of the masker. This result is more consistent with a larger role of the TFS in the target in that the representation of masker TFS likely becomes poorer as SNR (in the dips) becomes more favorable. Accordingly, the nature of the masker TFS may not be critical for speech recognition in noise. It is difficult, however, to rule out the potential influence of masker TFS on speech recognition in noise. Because most previous studies systematically removed TFS cues from the target and the masker simultaneously, the relative contribution of the target and masker TFS to speech recognition in noise remains unclear.

To our knowledge, only one study ([Apoux and Healy, 2010a](#)) has examined the relative contribution of the target and masker TFS to the unmasking of speech. In their main experiment, the authors measured the intelligibility of unprocessed and vocoded consonants in the presence of unprocessed maskers. The results showed that removing the original TFS from the target signal caused intelligibility to drop by ~ 13 and ~ 17 percentage points in speech-shaped noise (SSN) and time-reversed speech (TRS), respectively. In a supplementary experiment, [Apoux and Healy \(2010a\)](#) measured the intelligibility of vocoded consonants in the presence of unprocessed or vocoded maskers. The results of this second experiment indicated no influence of the masker TFS when the masker consisted of SSN and a small influence (7% points) when it consisted of TRS. Because [Apoux and Healy \(2010a\)](#) were mainly interested in the effect of the off- and on-frequency spectral components of a masker on speech intelligibility, they did not include a condition in which the masker was vocoded and the target was unprocessed. Therefore, although their data are consistent with a limited role of masker TFS, it is difficult to determine with certainty the potential influence of masker TFS on speech recognition in noise from these data. Accordingly, the purpose of the present study was to examine the effect of disrupting independently the TFS in the target signal and that in the masker to determine their relative contribution to the unmasking of speech.

As the detrimental effect of vocoder processing is larger when envelope fluctuations are present in the masker, it is seemingly important to use various types of maskers when investigating the relative importance of target and masker TFS. Previous studies typically compared the effect of vocoder processing in the presence of only two different types of maskers. Some studies compared steady to AM SSN maskers ([Gnansia et al., 2009](#); [Hopkins and Moore, 2009](#)), whereas others compared SSN to speech maskers ([Qin and Oxenham, 2006](#); [Apoux and Healy, 2010a](#)). Although the former comparison provides a direct assessment of the influence of AM in the masker, it provides limited, if any, insight into the role of masker TFS. In contrast, the latter comparison may provide some insight into the role of masker TFS, but because the maskers also differ in terms of their envelope, it is difficult to determine the relative influence of envelope and TFS. Accordingly, the present study included all three masker types. However, a recent study by [Rhebergen et al. \(2005\)](#) indicated that time-reversed and regular speech do not produce the same amount of masking. Therefore, the present study included both TRS and regular speech for a total of four different types of maskers.

II. METHOD

A. Subjects

Thirty-four NH listeners participated in the present experiment (32 females). Their ages ranged from 19 to 29 years (average = 21 years). All listeners had pure-tone air-conduction thresholds of 20 dB hearing level or better at octave frequencies from 250 to 8000 Hz (ANSI, S3.6-2004). 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.

B. Speech material and processing

The target stimuli consisted of 16 consonants (/p, t, k, b, d, g, θ , f, s, $\int$, δ , v, z, $\mathfrak{z}$, m, n/) in /a/-consonant-/a/ environment recorded by four speakers (two for each gender) for a total of 64 vowel-consonant-vowel utterances (VCVs; [Shannon et al., 1999](#)). The sampling rate for all recordings was 22 050 Hz. The target stimuli were either presented alone (Quiet) or with a masker. Four masker types were tested. 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). Another masker type, SMN, was created by modulating the broadband SSN with the broadband envelope obtained from a sentence randomly selected from the AzBio test ([Spahr et al., 2007](#)). Each envelope was obtained by half-wave rectification and low-pass filtering at 400 Hz (50-order FIR filter). The two other masker types were sentences randomly selected from the AzBio test. In one case, the sentences were left intact (SPE). In the other case, the sentences were played backward to eliminate to some extent linguistic content and limit confusions with the target, while preserving speech-like acoustic characteristics (TRS). Because the spectral shaping parameters for the SSN were selected to

approximate the long-term average spectrum of the AzBio sentences, the average long-term spectra of all four maskers matched.

Prior to combination, target and masker stimuli were filtered into 30 contiguous frequency bands ranging from 80 to 7563 Hz using two cascaded 12th-order digital Butterworth band-pass filters. Stimuli were filtered in both the forward and reverse directions so that the filtering produced zero phase distortion (for details, see Apoux and Healy, 2009). Each band was one ERB_N wide so that the filtering roughly simulated the frequency selectivity of the normal auditory system. Subjects were always presented with 30 bands. Target and masker bands were either left intact (UNP) or they were processed through a sine-wave vocoder to replace the original TFS with a tone in each band (VOC). Three VOC conditions were tested. In one condition, only the target bands were vocoded (VOC_t). In another condition, only the masker bands were vocoded (VOC_m). Finally, both target and masker bands were vocoded (VOC_{tm}). The vocoder processing was implemented as follows. The envelope was extracted from each band by half-wave rectification and low-pass filtering at cf_m (eighth-order Butterworth, 48 dB/octave roll-off). The value for cf_m was computed independently for each band so that it equaled half the bandwidth of the ERB_N -width filter centered at the sinusoidal carrier frequency (Apoux and Bacon, 2008). The filtered envelopes were then used to modulate sinusoids having frequencies equal to the center frequencies of the bands on an ERB_N scale and random starting phase. In the noise conditions, all 30 masker bands were presented simultaneously with the 30 target bands. The duration of the masker was always equal to that of the target speech.

The overall level of the 30 summed target speech bands was normalized and calibrated to be 65 dB(A) sound pressure level. The level of the 30 summed masker bands was adjusted to achieve 0 or -6 dB SNR when compared to the 30 summed target speech bands.

C. Procedure

Twenty-eight listeners were divided randomly into two groups. The remaining six listeners participated in both parts of the study, resulting in a total of 20 individual data sets (i.e., 20 “subjects”) per group. Each group was tested in one SNR condition: 0 or -6 dB. All combinations of VOC processing (UNP, VOC_t , VOC_m , and VOC_{tm}) and masker type (SSN, SMN, TRS, and SPE) were tested, resulting in 16 conditions. In addition, subjects were also tested in quiet with unprocessed and vocoded target speech, yielding a total of 18 conditions.

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 (Sennheiser Electronic Corporation, Old Lyme, CT). The experiments were controlled using custom MATLAB (The MathWorks, Natick, MA) routines running on PCs equipped with Echo Gina24 high-quality digital/analog converters (Echo Digital audio Corporation, Santa Barbara, CA). Percent correct identification was measured using a single-interval 16-alternative forced-choice procedure. Listeners

reported the perceived consonant by using the computer mouse to select 1 of 16 buttons on the computer screen. Approximately 30 min of practice was provided prior to data collection. Practice consisted of four blocks and all 64 VCVs were presented in each block in random order. Practice stimuli were processed through the sine-wave vocoder described earlier. Feedback was given during the practice session but not during the experimental session. After practice, each subject completed the 18 experimental blocks in random order. Each experimental block also corresponded to recognition of all 64 VCVs, presented in random order. The total duration of testing, including practice, was ~ 2 h.

III. RESULTS

The mean percent correct identification scores for the 0 and -6 dB SNR conditions are presented in Figs. 1 and 2, respectively. In both figures, each panel shows the data obtained in one of the four masker-type conditions as a function of the VOC processing. Although the VOC conditions do not form a continuum, they are connected with dotted lines to facilitate comparison. The standard deviation across listeners ranged from 4.5% to 7.5% points (mean = 5.9% points) in the 0 dB SNR condition and from 5.5% to 14% points (mean = 7.6% points) in the -6 dB SNR condition, with no notable difference across masker type or VOC processing condition. As expected, performance in quiet was very good ($>95\%$) and no difference was observed between the UNP and the VOC conditions (not displayed). Performance was systematically lower in the presence of the maskers. Also not surprisingly, performance was generally better at the higher SNR and/or when fluctuations were introduced in the masker. Overall, the effect of masker type and VOC processing was very similar across the two SNRs.

A repeated-measures analysis of variance with factors masker type and VOC processing was performed for each SNR. There was a main effect of masker type at 0 dB SNR [$F(3,57) = 166, p < 0.001$] and at -6 dB SNR [$F(3,57) = 522, p < 0.001$]. The effect of VOC processing was also significant at both 0 dB [$F(3,57) = 40.5, p < 0.001$] and -6 dB SNR [$F(3,57) = 28.6, p < 0.001$]. The two-way interactions were also all significant ($p < 0.001$), suggesting a differential role of TFS in various backgrounds. Multiple pairwise comparisons (corrected paired t -tests)² were performed separately for each SNR condition and significant contrasts are indicated by double-headed arrows in Figs. 1 and 2. Because the pattern of results was very similar for the two SNRs, the data will not be discussed separately in what follows.

It is apparent from Figs. 1 and 2 that two clear patterns of results were obtained. The first pattern was only observed for the SSN masker. For this masker, vocoding the masker (VOC_m) or the entire sound mixture (VOC_{tm}) did not have a significant effect on consonant recognition. This latter result is consistent with that observed in previous work that indicated a limited role for the TFS of steady noise. However, performance decreased significantly when only the target was vocoded (VOC_t). The second pattern of results was observed for all the other masker type conditions. Vocoding the masker (VOC_m) did not have a significant effect on

FIG. 1. Averaged percent correct scores for consonant identification as a function of VOC processing [unprocessed (UNP), target only (VOC_t), masker only (VOC_m) and entire stimulus (VOC_{tm})]. Each panel corresponds to one of the four masker types. Arrows indicate a significant difference between two conditions. Errors bars indicate 1 standard deviation. The signal-to-noise ratio was 0 dB.

consonant recognition in five of the six conditions, whereas a significant effect of vocoding the target (VOC_t) was systematically observed. In contrast to the results with the SSN masker, a significant effect of vocoding the mixture was observed with all three masker types and for both SNR conditions. Again, this is consistent with previous work that has typically involved removal of TFS from the entire sound mixture and has demonstrated reduced performance in AM masker conditions.³ Finally, a significant difference between the performance in the VOC_m condition and that in the VOC_{tm} condition was observed for all of the non-SSN masker conditions.

The percentage transmission of features voicing, manner, and place of articulation computed for each experimental condition from the averaged consonant confusion

matrices is shown in Figs. 3–5, respectively (Miller and Nicely, 1955). In each figure, the top panel is for 0 dB SNR and the bottom panel is for –6 dB SNR. To better illustrate the effects of VOC processing, data are presented in terms of information transmitted relative to the UNP condition (in percentage points). Accordingly, negative values indicate a drop in the transmission of the feature. The effect of VOC processing on the transmission of voicing was reasonably consistent across SNRs (Fig. 3). There were two set of conditions for which a clear detrimental effect of VOC processing was observed. The first set was for the VOC_t condition, but it was limited to those conditions where the background was either a SSN or a SMN. The second set was for the VOC_{tm} condition, especially when the masker was modulated. The effect of VOC processing on the transmission of

FIG. 2. Same as Fig. 1 except that the signal-to-noise ratio was –6 dB.

FIG. 3. Percent of information transmitted for the feature voicing relative to the unprocessed condition for the four masker types. The upper and lower panels correspond to the 0 and -6 dB SNR conditions, respectively.

manner was also fairly consistent across SNRs with one notable exception (Fig. 4). Transmission of manner generally decreased for all masker-type conditions whenever TFS information was removed from the target. At -6 dB SNR, however, no considerable decrease was observed for the SSN condition. Similar to what was observed for voicing, vocoding the entire stimulus reduced the transmission of manner principally when the masker was modulated. Although the general pattern was comparable (Fig. 5), the effect of VOC processing on the transmission of place was larger for 0 dB SNR than for -6 dB SNR. Overall, an effect of removing TFS information from the target was observed with all masker types, but this effect was limited for the -6 dB SNR condition. Again, vocoding the entire stimulus produced a drop in information transmission but only when the masker was modulated. Another common result across these three features is that transmission was generally unaffected by the disruption of the masker TFS only (VOC_m), irrespective of the masker type.

IV. DISCUSSION

Recently, it has been suggested that changes in phase locking of auditory nerve discharges when a valley occurs may be used by the auditory system to determine that a signal is present in the dip (Lorenzi *et al.*, 2006). This

FIG. 4. As Fig. 3, but for the feature manner of articulation.

hypothesis is consistent with the finding that speech recognition in a fluctuating background decreases when the TFS of the sound mixture is replaced with noise or tones. Indeed, the target and masker TFS are no longer “disparate” after such manipulation. Based on this hypothesis, one could reasonably assume that two conditions having identical disparity or distance between target and masker TFS should produce comparable intelligibility. The results obtained with the speech maskers (SPE and TRS), however, were not consistent with this prediction. Consider for instance the SPE masker. In the VOC_t condition, the target TFS consisted of a series of tones and the masker TFS consisted of the TFS of speech. Inversely, in the VOC_m condition the target TFS consisted of the TFS of speech and the masker TFS consisted of a series of tones. If Δ_{tfs} represents the distance between the target and the masker TFS in the VOC_t condition, one may reasonably assume Δ_{tfs} to also represent the distance between the target and the masker TFS in the VOC_m condition. Based on the disparity hypothesis, the same effect should have been observed in the VOC_t and VOC_m conditions. The present results, however, suggest otherwise. Although some indefinite factors may have obscured the influence of disparity, the results of the present study do not support the hypothesis of a mechanism based on disparities between the target and masker TFS.

As always, it is challenging to determine the extent to which results obtained with a limited set of consonants might generalize to sentence recognition or connected speech.

FIG. 5. As Fig. 3, but for the feature place of articulation.

However, it is difficult to imagine that the factors typically involved in sentence recognition, but not consonant recognition (i.e., the linguistic and semantic cues), will interact with TFS information in such a way that a different pattern of results will be observed. In contrast, the use of sentences as targets may increase the masking power of the SPE masker. Indeed, it has been reported that “incomprehensible” speech maskers (i.e., speech maskers that do not convey a meaning) are less detrimental to intelligibility than comprehensible ones (Rhebergen *et al.*, 2005). In the present study, no difference was observed between the TRS (incomprehensible) and the SPE (comprehensible) maskers. However, only a short portion of the masker sentences was presented to match the brief duration of the target stimuli (a few hundred milliseconds). Presumably, no meaning could be extracted from such brief sentence segments. As a consequence, SPE maskers did not provide additional masking compared to the TRS maskers. Increasing the duration of the target stimuli, which in effect increases the duration of the masker, will presumably result in increased informational masking (e.g., Kidd *et al.*, 2008).

Comparison across masker type provides further evidence for the limited role of the masker TFS in the unmasking process. To better illustrate the effect of masker type, averaged data from Figs. 1 and 2 have been replotted in Fig. 6. Each panel shows the four averaged functions obtained at a given SNR. Of particular interest here is the

FIG. 6. Averaged percent correct scores for consonant identification as a function of VOC processing [unprocessed (UNP), target only (VOC_t), masker only (VOC_m) and entire stimulus (VOC_{tm})] for the four masker types. The upper and lower panels correspond to the 0 and -6 dB SNR conditions, respectively.

comparison of performance in the SMN and SPE conditions. Although the same pattern of results was observed in these conditions, recognition scores in the SMN condition were substantially lower than for the SPE condition. Although one may be tempted to attribute this lower performance to differences in the TFS of these maskers, it is apparent from Fig. 6, especially for -6 dB SNR, that the drop in performance was fairly similar across VOC processing conditions. Because the TFS of the maskers was replaced with tones in some of these processing conditions, the difference between SMN and SPE cannot be attributed to the nature of their TFS. More importantly, the finding that the difference between SMN and SPE maskers remained constant across VOC processing conditions suggests a limited influence of the masker TFS.

A question that emerges at this point is: why was the masking produced by the SMN masker higher than that produced by the speech maskers if the nature of the masker TFS does not play a role in speech recognition? The SMN masker and the speech maskers differed markedly in their TFS. However, these maskers also differed in their temporal envelopes. As mentioned in Sec. II, SMN maskers were created by imposing the broadband envelope of a sentence on a speech-shaped noise. As a result, amplitude fluctuations were highly correlated across the spectrum. In contrast, it is well established that narrowband speech envelopes are only partially correlated across frequency and that this correlation

FIG. 7. Narrowband envelopes derived from a speech-modulated noise (solid lines) and a sentence (dashed lines). The panels from top to bottom show the envelopes derived from bands 5, 10, 15, and 20, respectively (center frequency = 272, 630, 1 241, and 2 289 Hz). These envelopes were obtained using the techniques described in the Method section.

decreases with increasing rate and/or spectral separation (e.g., Crouzet and Ainsworth, 2001; Healy *et al.*, 2005; Apoux and Bacon, 2008).

One consequence of the limited correlation between narrowband speech envelopes is an increased probability to observe frequency regions dominated by the target signal at any moment in time. This interpretation seems to contradict the results of Howard-Jones and Rosen (1993) showing that speech recognition in noise is better when the modulations of the masker are correlated across the spectrum. As noted by Cooke (2003), however, it is not clear whether similar results would be found for structured maskers such as speech. More importantly, Howard-Jones and Rosen (1993) only used noise as a masker and the overall percentage of “clean” speech was kept constant across conditions (i.e., 50%). In the present experiments, it may be assumed that the overall percentage of clean speech was notably larger in the presence of the speech masker than in the presence of the modulated noise. This larger proportion of clean speech may have compensated for the potential drop in performance that could have resulted from the introduction of unmodulated glimpses. To illustrate this idea, four narrowband envelopes extracted from a SMN masker (solid lines) and the corresponding narrowband envelopes from the original sentence (dashed lines) are plotted in Fig. 7. The envelopes were obtained as described in Sec. II. Figure 7 illustrates two important differences between the two masker types. First, the envelope correlation across frequency appears higher for the SMN masker than for the SPE masker, especially for the lower-rate AMs. Second, the number of glimpsing opportunities is generally larger with the SPE masker, except in the lowest bands. As speech information is primarily located in the 400–2500 Hz frequency region for these target stimuli (Apoux and Healy, 2010b), it is not surprising that sentences provided less masking than speech-modulated noise. The lower amount of masking produced by the speech maskers can therefore be attributed to these additional glimpsing opportunities.

Overall, the effect of masker type can be attributed almost exclusively to differences in envelope fluctuations.

When the envelope of the masker was the most constant (SNN), fewer glimpsing opportunities existed and the amount of masking was maximal. When speech AM was introduced in the masker (SMN), the amount of masking was substantially reduced, especially at the lower SNR. This masking release (MR; e.g., Fullgrabe *et al.*, 2006) was further increased when using natural speech as a masker because natural speech presumably offers more glimpsing opportunities.

Considering that the primary purpose of this study was to assess the relative contribution of the target and masker TFS to the unmasking of speech, it is not surprising that the present data do not clearly establish whether the auditory system uses TFS cues to determine which portions of a sound mixture have a relatively favorable SNR or to directly identify speech sounds. However, it is worth speculating whether one of these two mechanisms was predominantly involved in the present results. For the SSN masker, the fact that a significant effect of removing TFS information from the target (VOC_t) was observed, whereas no effect was observed in the VOC_{tm} condition may be interpreted as evidence that subjects primarily used TFS cues to segregate speech from noise (as opposed to identifying speech sounds). Because TFS information was removed from the target in both conditions, it is indeed reasonable to assume that the decrease in performance observed in the VOC_t condition was not caused by the missing speech information. Instead, one may assume that the ability to segregate speech from noise was impaired when only the TFS of the target was disrupted, but this ability remained unaffected (i.e., similar to that in the UNP condition) when the disruption was applied to the entire stimulus. A possible interpretation has to do with the reduced fluctuations in the envelope of the SSN masker compared to the other maskers. Indeed, the physical envelope of the SSN masker was steadier than that of the other maskers. As a consequence, the fluctuations in the sound mixture, though reduced in depth, primarily conveyed information about the target, therefore facilitating the segregation of the target from the masker. This interpretation is consistent with the results observed with the other masker

types as the presence of more extensive envelope fluctuations in the maskers presumably made the task of segregating speech and masker envelopes more difficult in VOC_{tm}. In these cases, only some fluctuations pertained to the target, whereas the others pertained to the masker.

As mentioned earlier, most previous studies reported a larger effect of disrupting TFS information on speech recognition in the presence of a fluctuating background. This differential effect predictably results in reduced MR when TFS information is disrupted. To assess the effect of selectively disrupting the TFS of the target or that of the masker on MR, the difference between recognition scores in the SSN and SMN conditions was calculated for each processing condition (see also Fig. 6). At 0 dB SNR, MR was ~12% points in the UNP condition. Removing the TFS from the target (VOC_t) did not have a large effect on MR (14% point difference). In contrast, MR dropped to ~7% points when the TFS was only removed from the masker (VOC_m). However, this smaller difference was mainly due to an increase in performance in the SSN condition as performance in the VOC_m condition remained similar to that in the UNP condition with the modulated masker. No MR was observed when the TFS of the entire sound mixture was replaced with tones. In the -6 dB SNR condition, MR was also comparable in the UNP, VOC_t, and VOC_m conditions (28%–32% points). Although reduced to 16% points, MR was still observed in the VOC_{tm} condition. Overall, the present data are consistent with previous work showing that MR is substantially reduced in the absence of TFS cues. However, the reduced MR could not be attributed to a disruption of either the target or the masker TFS.

V. SUMMARY AND CONCLUSIONS

The primary goal of the present study was to assess the relative contribution of the target and masker TFS to speech recognition in noise. Based on the two mechanisms suggested by previous work and a limited set of data from Apoux and Healy (2010a), it was hypothesized that the TFS of the masker does not play a critical role in the unmasking of speech. To test this hypothesis, the TFS of the target speech and that of the masker were manipulated independently to assess their relative effect on speech intelligibility. Out of the eight possible comparisons between unprocessed and vocoded masker, only one was significant (0 dB SNR, TRS). In contrast, all eight possible comparisons between unprocessed and vocoded target were significant. Such clear contrast between the effect of removing TFS information from the target and that of removing TFS information from the masker is largely consistent with the above hypothesis. In other words, the normal auditory system relies primarily on the TFS in the target to detect the portions of a sound mixture with relatively favorable SNR and/or to determine the identity of speech sounds. More generally, the present data indicate that the unmasking of speech is primarily influenced by the nature of the target TFS.

ACKNOWLEDGMENTS

This research was supported by grants from the National Institute on Deafness and Other Communication Disorders

(NIDCD Grant No. DC009892 awarded to F.A. and DC008594 awarded to E.W.H.). The assistance of Carla Y. Berg and Sarah E. Yoho in collecting and analyzing the data is gratefully acknowledged. We thank Brian C. J. Moore and two anonymous reviewers for their helpful comments during the review process.

¹Gilbert and Lorenzi (2006) and Lorenzi *et al.* (2006) used the Hilbert transform to isolate TFS from the envelope, whereas Sheft *et al.* (2008) derived the analytic signals from the output of a bank of 2-ERB wide analysis filters and estimated a phase modulation function or a FM function from the output of each analysis filter. See cited studies for details.

²Paired *t*-test results were corrected using the incremental application of Bonferroni correction described by Benjamini and Hochberg (1995).

³The fact that TFS degradation had a modest effect on consonant recognition relative to that observed in previous studies may be explained by the large number of processing bands used in the present study.

- ANSI S3.6-2004 (2004). "Specifications for audiometers" (American National Standards Institute, New York).
- Apoux, F., and Bacon, S. P. (2008). "Differential contribution of envelope fluctuations across frequency to consonant identification in quiet," *J. Acoust. Soc. Am.* **123**, 2792–2800.
- 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. (2010a). "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. (2010b). "Auditory channel weights for consonant recognition in normal-hearing listeners," *J. Acoust. Soc. Am.* **127**, 1191(A).
- 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.
- Benjamini, Y., and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300.
- 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 Acoustic Cues for Sound Analysis, Aalborg, Denmark, September 2001.
- Cooke, M. (2003). "Glimpsing speech," *J. Phonetics* **31**, 579–584.
- Fullgrabe, 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., Jourdes, V., and Lorenzi, C. (2008). "Effect of masker modulation depth on speech masking release," *Hear. Res.* **239**, 60–68.
- Gnansia, D., Pean, 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.
- Healy, E. W., Kannabiran, A., and Bacon, S. P. (2005). "An across-frequency processing deficit in listeners with hearing impairment is supported by acoustic correlation," *J. Speech Lang. Hear. Res.* **48**, 1236–1242.
- 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.
- Howard-Jones, P. A., and Rosen, S. (1993). "Unmodulated glimpsing in 'checkerboard' noise," *J. Acoust. Soc. Am.* **93**, 2915–2922.
- Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (2008). "Informational masking," in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer Science + Business Media, LLC, New York), pp. 143–190.

- Lorenzi, C., Gilbert, G., Carn, 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.
- Miller, G. A., and Nicely, P. E. (1955). "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.* **27**, 338–352.
- 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.
- Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2005). "Release from informational masking by time reversal of native and non-native interfering speech," *J. Acoust. Soc. Am.* **118**, 1274–1277.
- Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Shannon, R. V., Jansvold, A., Padilla, M., Robert, M. E., and Wang, X. (1999). "Consonant recordings for speech testing," *J. Acoust. Soc. Am.* **106**, L71–L74.
- Spahr, A. J., Dorman, M. F., and Loiselle, L. H. (2007). "Performance of patients using different cochlear implant systems: effects of input dynamic range," *Ear Hear.* **28**, 260–275.
- Sheft, S., Ardoint, M., and Lorenzi, C. (2008). "Speech identification based on temporal fine structure cues," *J. Acoust. Soc. Am.* **124**, 562–575.
- 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.