Hearing Research 255 (2009) 99–108

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research papers

On the number of auditory filter outputs needed to understand speech: Further evidence for auditory channel independence ☆

Frédéric Apoux *, Eric W. Healy

Department of Speech and Hearing Science, The Ohio State University, OH, USA

ARTICLE INFO

Article history:
Received 23 January 2009
Received in revised form 1 June 2009
Accepted 10 June 2009
Available online 16 June 2009

Keywords:
Peripheral filtering
Speech recognition
Glimpsing
Auditory channel independence

ABSTRACT

The number of auditory filter outputs required to identify phonemes was estimated in two experiments. Stimuli were divided into 30 contiguous equivalent rectangular bandwidths (ERB_N) spanning 80–7563 Hz. Normal-hearing listeners were presented with limited numbers of bands having frequency locations determined randomly from trial to trial to provide a general view, i.e., irrespective of specific band location, of the number of 1-ERB_N-wide speech bands needed to identify phonemes. The first experiment demonstrated that 20 such bands are required to accurately identify vowels, and 16 are required to identify consonants. In the second experiment, speech-shaped noise or time-reversed speech was introduced to the non-speech bands at various signal-to-noise ratios. Considerably elevated noise levels were necessary to substantially affect phoneme recognition, confirming a high degree of channel independence in the auditory system. The independence observed between auditory filter outputs supports current views of speech recognition in noise in which listeners extract and combine pieces of information randomly distributed both in time and frequency. These findings also suggest that the ability to partition incoming sounds into a large number of narrow bands, an ability often lost in cases of hearing impairment or cochlear implantation, is critical for speech recognition in noise.

Published by Elsevier B.V.

1. Introduction

Most natural sounds, such as speech, are highly modulated both in time and frequency. As a consequence, the relationship between speech and noise intensities, i.e., the signal-to-noise ratio (SNR) is usually non-uniform across frequency and may change rapidly over brief periods of time. Because sounds produced by independent sources are not likely to be correlated, there is often a fair probability to observe frequency regions dominated by the signal of interest – the target signal – at any moment in time. Furthermore, acoustic speech cues are highly redundant in the frequency domain in that individual cues can be degraded to various degrees without affecting overall recognition (e.g., Shannon et al., 1995; Warren et al., 1995). Accordingly, the combination of even a restricted number of frequency regions in which the target signal is relatively preserved from the background may be sufficient to reconstruct an interpretable representation of the target speech signal.

Consistent with the above, current models of speech recognition in noise suggest that the auditory system takes advantage of this probability to observe frequency regions dominated by the target speech signal. One model recently proposed by Cooke (2003, 2005, 2006) integrates and extends previous views of speech recognition in noise. In accord with the view initially suggested by Miller and Licklider (1950), this model assumes that normal-hearing (NH) listeners take advantage of temporal gaps present in the masker. The Miller and Licklider "listening-in-the-dips" hypothesis was primarily motivated by the fact that intelligibility in noise increases substantially when spectral and/or temporal gaps are introduced in the masker (Miller and Licklider, 1950; Festen and Plomp, 1990; Howard-Jones and Rosen, 1993; Gustafsson and Arlinger, 1994; Buss et al., 2003; Nelson et al., 2003; Füllgrabe et al., 2006; Iyer et al., 2007). There were, however, several limitations to the listening-in-the-dips hypothesis. In order to circumvent these limitations, two major amendments were proposed by Cooke. First, it is now suggested that speech recognition in noise does not rely solely on momentary improvements in overall (across frequency) SNR but more generally on the ability to extract speech information (i.e., acoustic speech cues) from time–frequency regions that contain a relatively undistorted view of local signal properties. This first

☆ Portions of this work were presented in "Phoneme recognition as a function of the number of auditory filter outputs", Proceedings of the ASA/EAA/SFA joint conference "Acoustics'08" Paris.

* Corresponding author. Address: Department of Speech and Hearing Science, Speech Psychoacoustics Laboratory, The Ohio State University, 110 Pressey Hall, 1070 Carmack Rd., Columbus, OH 43210, USA. Tel.: +1 614 292 8059.

E-mail addresses: fred.apoux@gmail.com (F. Apoux), healy.66@osu.edu (E.W. Healy).

0378-5955/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.heares.2009.06.005

amendment was introduced to account for those situations where listeners maintain a communication while there are no gaps in the background (steady backgrounds, such as speech-shaped noise or multitalker babble). Second, while it is assumed that the normal auditory system extracts cues primarily from frequency regions containing the clearest available views of the speech signal (i.e., the most favorable SNRs available), it is further suggested that weak elements of speech lying below the noise level may also contribute to overall intelligibility. The contribution of acoustic speech cues extracted from frequency regions containing substantial amounts of noise is consistent with the results of several behavioral studies demonstrating that speech recognition in noise does not rely solely on time–frequency regions entirely dominated by the target speech signal (Drullman, 1995; Brungart et al., 2006; Li and Loizou, 2008). While the concept of glimpse was first introduced by Miller and Licklider (1950), glimpses will refer here to time–frequency regions that contain a relatively undistorted view of local signal properties, and the model proposed by Cooke will be referred to as the glimpsing model.

The glimpsing model is also largely inspired by psychophysical studies on pure-tone masking and the concept of critical bands (Fletcher, 1940). As pointed out by Celmer and Bienvenue (1987), the concept of critical bands implies that the peripheral auditory system acts as a kind of noise reduction system. In this view, the peripheral auditory system partitions incoming sounds into a series of bands, the critical bands, and the bands containing a large amount of noise – or considered as noise – are simply ignored. It should be noted that the degree of frequency resolution that can be achieved by the peripheral auditory system is critical: the narrower the bandwidth of the auditory filters, the more noise the system can reject. Accordingly, the glimpsing model assumes that the "auditory periphery decomposes the auditory mixture to [...] time–frequency units [...] with the size of each unit representing the smallest auditory event that can be resolved" (Li and Loizou, 2008), and that the internal representation of the target signal is reconstructed by combining those time–frequency units considered to pertain to the target signal (i.e., the bands containing a large amount of noise are ignored). It is apparent then that the internal representation of a target signal in noise necessarily results from the combination of a subset of auditory filter outputs.

A fundamental question that emerges at this point is how many of these auditory filter outputs – within the range of frequencies relevant for speech recognition – are necessary to understand speech. It might not be possible, however, to estimate directly the relationship between speech intelligibility and number of auditory filter outputs. Indeed, auditory filters greatly overlap in the frequency domain (e.g., Fletcher, 1940), making it impracticable to excite a single auditory filter without stimulating adjacent filters. In this study, a common simplifying assumption was employed, under which peripheral filtering is represented by a series of contiguous bands each roughly corresponding to the width of a critical band. Equivalent rectangular bandwidth (ERB_N; Glasberg and Moore, 1990), a widely accepted functional estimate of auditory filter bandwidth, was chosen. Subjects were then presented with n speech bands selected from the possible auditory filter width bands. We reasoned that the n auditory filters centered on the n target speech bands would be primarily excited by energy from the target signal. Therefore, it may reasonably be assumed that *no fewer* than n auditory filter outputs containing unmasked speech should be available to the subjects.

Potentially, methods developed for predicting speech intelligibility under a variety of noise conditions such as the Articulation Index (AI; French and Steinberg, 1947; Kryter, 1962; ANSI S3.5-1969) and the subsequent Speech Intelligibility Index (SII; ANSI S3.5-1997) offer a way to assess the number of auditory filter width bands necessary to understand speech. As mentioned earlier, however, speech recognition in noise most likely relies on the combination of a limited number of auditory filter outputs, which may be sparsely distributed across the spectrum. A limitation of the AI approach is that it was not designed for predicting the intelligibility of spectrally disjoint frequency bands, and therefore, these conditions fall outside its scope. It is not surprising then that the AI cannot account for the synergistic and redundant interactions among the various spectral regions of the speech spectrum. More specifically, the calculation procedure does not incorporate combination rules other than addition across bands, and therefore has difficulty dealing with these considerable interactions (Breeuwer and Plomp, 1984, 1985, 1986; Warren et al., 1995; Lippman, 1996; Healy and Warren, 2003). As a result, intelligibility often exceeds the predictions made by AI/SII models when the audible speech spectrum is partitioned into two or more spectrally disjoint frequency bands (Kryter, 1962; Grant and Braida, 1991). Another limitation involves the inability to accurately predict the intelligibility of speech in the presence of non-stationary maskers (e.g., Dubno et al., 2002). A direct measure of the number of auditory filter width bands necessary to understand speech in noise is therefore needed.

Two experiments were designed to investigate the relationship between number of available 1-ERB_N width speech bands and intelligibility. In both experiments, the location of the speech bands was randomized from trial to trial. The randomization served two purposes. First, it is well established that speech information is not distributed uniformly across frequency. As a result, the contribution of a speech band varies with its spectral location. Moreover, synergistic interactions between bands also affect the total amount of information present. For example, it has been demonstrated that the intelligibility of discrete 1/3-octave speech bands varies as a function of the spectral separation between these bands (Healy and Warren, 2003). It was therefore important to minimize the effects of spectral location so that the results would provide a general view, i.e., irrespective of band location, of the number of auditory filter width speech bands needed to understand speech. This was achieved by randomizing the spectral location of the bands. A second purpose for band randomization involves the fact that speech energy varies constantly in both time and frequency. While in some instances it may exhibit regularities, background noise also typically fluctuates randomly in both time and frequency. As a consequence, listeners cannot predict which frequency region will convey usable speech information and cannot know which auditory channel to attend. The randomization of speech band location replicated to some extent this moment-to-moment unpredictability.

In the first experiment, subjects were asked to identify phonemes when presented with a subset of 1-ERB_N width speech bands. To allow each speech band to provide its maximal contribution without the potentially interfering influence of noise, no noise was added. Indeed, the presence of noise in a target speech band would inevitably affect the contribution of that band to overall intelligibility. By keeping the SNR high, the potential contribution or "availability" of each band remained at a maximum. Therefore, when n bands were presented, listeners had the opportunity to extract as much information as possible from any of the n bands. In the second experiment, noise was presented simultaneously with the target speech. The target and masker bands were spectrally interleaved so that overlap in the spectral domain (i.e., peripheral masking) was limited and again, the speech bands remained as undistorted as possible. An advantage of the interleaving arrangement is that listeners are forced to select a limited number of auditory filter outputs and ignore the others, thus simulating to some extent speech recognition in noise as advocated by glimpsing models. A second advantage of the interleaving arrangement is that it

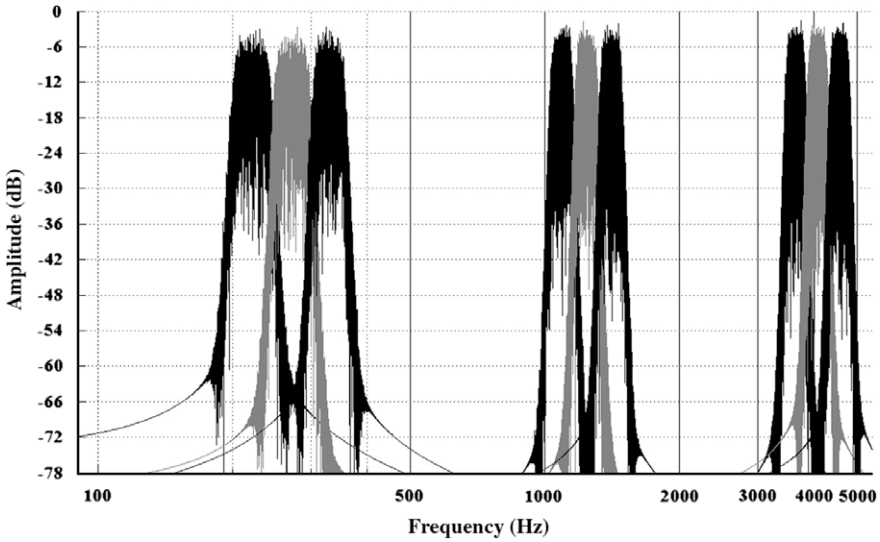

Fig. 1. Long-term average spectra of the output of filter numbers 4, 5, 6, 14, 15, 16, 24, 25, and 26. Their respective center frequencies were 221, 272, 329, 1091, 1241, 1408, 3642, 4082 and 4572 Hz, and each was 1-ERB_N wide. The input signal was a 60 s white noise.

should limit the contribution of frequencies lying outside the nominal bandwidth. Indeed, it was anticipated that subjects might be able to use off-frequency cues from adjacent non-speech bands or listen in the transition bands (e.g., Healy, 1998; Warren et al., 2004). Finally, we reasoned that the comparison between intelligibility in quiet and in spectrally interleaved noise may provide important indications of the contribution to overall intelligibility of auditory filters not centered on the speech band.

2. Materials and methods

2.1. Subjects

Twenty-four NH subjects participated. Their ages ranged from 22 to 37 years (average = 23.5 years). Normal hearing was defined as pure-tone air-conduction thresholds of 20 dB HL or better (ANSI, S3.6-2004) for octave frequencies from 125 to 8000 Hz. All participants were native speakers of American English and received course credit for their participation. This study was approved by the University Institutional Review Board.

2.2. Speech material and processing

The target stimuli consisted of 9 vowels (/æ, ɔ, ɛ, i, ɪ, ɑ, u, ʊ, ʌ/) in /h/-vowel-/d/ environment recorded by six speakers (three for each gender) for a total of 54 consonant–vowel–consonant utterances (CVCs; Hillenbrand et al., 1995), and 16 consonants (/p, t, k, b, d, g, ɵ, f, s, ʃ, ð, v, z, ʒ, 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 background noise was a simplified speech spectrum-shaped noise (SSN; constant spectrum level below 800 Hz and 6 dB/oct roll-off above 800 Hz) or a sentence randomly selected from the speech perception in noise test (Kalikow and Stevens, 1977). All sentences were played backward to eliminate to some extent linguistic content (see, Rhebergen et al., 2005) and limit confusions with the target while preserving their speech-like acoustic characteristics (reversed speech; RS). The duration of the masker was always equal to target speech duration.

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 filters. Stimuli were filtered in both the forward and reverse directions (i.e., zero phase digital filtering) so that the filtering process would produce zero phase distortion.[1] Each band was one ERB_N wide so that the filtering simulated the frequency selectivity of the normal auditory system. To illustrate the amount of overlap between bands, the output of nine filters located in the low-, mid-, and high-frequency region in response to a 60 s white noise is shown in Fig. 1. In the conditions in which masker bands were present, n 1-ERB_N-wide speech bands were randomly selected among the 30 possible bands, and the remaining 30 minus n bands were replaced (i.e., filled) with masker bands. As a result, each one of the 30 possible bands contained either a target or a masker band and the manner in which the target and masker bands were arranged spectrally was random. While it generally implies some regularity, interleaved will refer to this random spectral arrangement of target and masker bands throughout this paper.

The target speech was normalized and calibrated so that its overall A-weighted output level was 65 dB when presented alone in the 30-band, i.e., "broadband", condition. The overall level of the 30 summed masker bands was adjusted to achieve a specific SNR when compared to the broadband target. Target and masker bands were combined after level adjustment so that the root mean square energy of any given band remained equal across conditions, irrespective of the total number of speech or noise bands. Because spectrum levels were held constant, overall levels of the stimuli generally increased with increased numbers of bands (Experiment 1) and overall SNR generally decreased with increased numbers of noise bands (Experiment 2). This approach was chosen because it best mimics what occurs in natural listening. Indeed, when portions of a target speech signal are masked by noise, the level of the other portions remains the same.

[1] The filters were designed using the function "butter" in Matlab. The order, N, of the digital Butterworth band-pass filters was initially set to 3. Because "butter" returns an order 2N filter when two cutoff frequencies are specified (i.e., band-pass filtering), the actual order of the digital Butterworth band-pass filters was 6. The implementation of zero-phase digital filtering in Matlab ("filtfilt" function), by processing the signal in both the forward and reverse directions, also doubled the order of the filter, yielding an effective order of 12. Finally, the filter order doubled again as the original stimuli were passed twice through these 12th-order digital Butterworth band-pass filters, resulting in an overall filter order of 24. As can be seen in Fig. 1, this filtering process yielded a slope that exceeded 6 dB/oct/order.

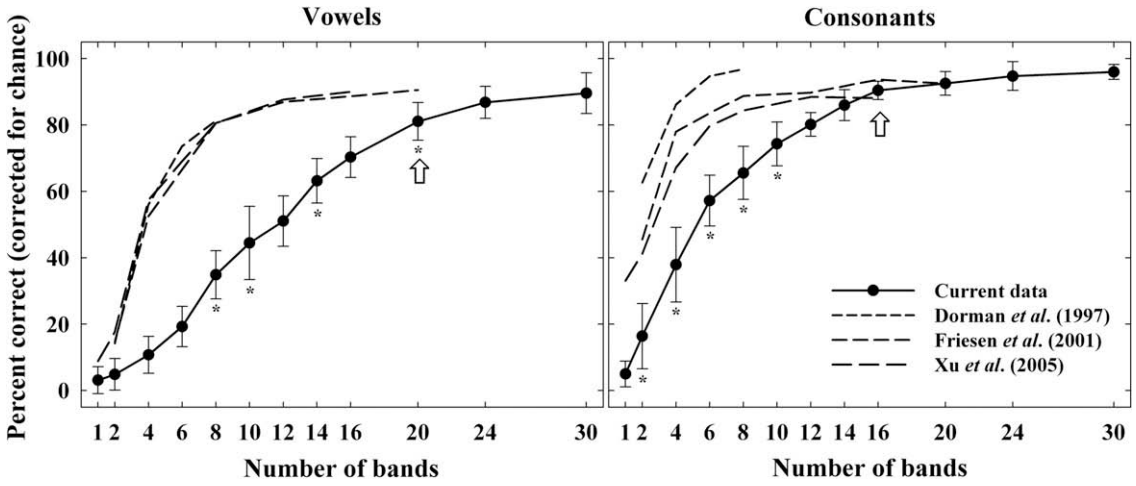

Fig. 2. Percent correct scores, corrected for chance, for vowel (left panel) and consonant (right panel) recognition as a function of the number of bands. The symbols represent data from the present study, in which ERB_N speech bands were employed. Dashed lines represent data from three previous studies using similar speech materials but vocoder processing. In each panel, the asymptotic performance is indicated by an arrow. Asterisks indicate conditions in which phoneme recognition was significantly different from that in the number-of-bands condition immediately to the left (lower number of bands). The errors bars show one standard deviation.

2.3. Procedure

Listeners were tested individually in a single-walled, sound-attenuated booth. Stimuli were played to the listeners binaurally through Sennheiser HD 250 Linear II circumaural headphones. The experiments were controlled using custom Matlab routines running on a PC equipped with high-quality D/A converters (Echo Gina24). Percent correct identification was measured using a single-interval, 9- or 16-alternative forced-choice procedure for the vowel and consonant tests, respectively. Listeners were instructed to report the perceived vowel or consonant and responded using the computer mouse to select 1 of 9 or 16 buttons on the computer screen. Prior to data collection, listeners twice completed recognition of all 54 CVCs or 64 VCVs in quiet with all 30 speech bands present. In what follows, a block will refer to recognition of all 54 CVCs or 64 VCVs. In each block, the order of the stimuli was always randomized. Visual on-screen feedback was provided after each trial during the practice session but not during the experimental sessions.

Two experiments were conducted. First, phoneme recognition was measured in quiet (Experiment 1). Twelve number-of-bands conditions were tested. In each condition, subjects were presented with n 1-ERB_N-wide speech bands (n = 1, 2, 4, 6, 8, 10, 12, 14, 16, 20, 24 or 30) selected randomly from trial to trial among the 30 possible bands. As a result, speech information was randomly distributed across frequency and this distribution varied for each phoneme presentation. No signal was presented in the non-speech bands. Twelve subjects were asked to identify vowels and the remaining 12 subjects were asked to identify consonants. Thus, each subject completed 12 blocks (648 or 768 trials) with each block corresponding to a number-of-bands condition. Then, phoneme recognition was measured in the same subjects in the presence of a simultaneous masker (Experiment 2). Like in Experiment 1, n 1-ERB_N-wide speech bands (n = 4, 8, 12, 16 or 24) were randomly selected on each trial among the 30 possible bands. In these noise conditions, however, 1-ERB_N-wide bands of SSN or RS were presented in the non-speech bands. All subjects performed the task with the same set of stimuli as in Experiment 1 (CVCs or VCVs). However, the subjects in each group were divided and assigned to one of the two background conditions (6 to SSN and 6 to RS). Six SNRs were employed. The SNR ranged from −12 to 18 dB and from −18 to 12 dB in 6-dB steps for SSN and RS, respectively. As a result, each subject completed 30 blocks (1620 or 1920 trials) corresponding to all combinations of number of target bands and SNRs for a given target/masker combination. In both experiments, blocks were presented in random order to avoid order effects.

3. Results

3.1. Experiment 1: Vowel and consonant recognition in quiet

Fig. 2 shows the percentage of vowels (left panel) and consonants (right panel) correctly identified as a function of the number of bands (symbols). Data from three vocoder studies using similar speech material (Dorman et al., 1997; Friesen et al., 2001; Xu et al., 2005) are also shown in each panel as dashed lines. These data are discussed later. For comparison, all scores were corrected for chance, using the appropriate chance level (e.g., Baskent and Shannon, 2006). With one band only, performance was at or only slightly above chance level (0% here). It then increased gradually with increasing numbers of bands. The increase was more gradual for vowels than for consonants (i.e., more bands were needed for the former to achieve a given performance), suggesting a greater importance of spectral cues in vowel recognition. For both sets of stimuli, highest performance was reached in the 30-band condition.

Separate one-way analyses of variance (ANOVA) with repeated measures were performed for each set of phonemes. The results indicated a main effect of number of ERB_N bands for vowels [$F(11, 143) = 232.89$, $p < 0.001$] and for consonants [$F(11, 143) = 307.65$, $p < 0.001$]. A post hoc test according to Tukey was used for all pairwise comparisons (alpha = 0.05). The main results of the post hoc tests are summarized in Fig. 2 as follows. Percent correct scores that were significantly different from those in the number-of-bands condition immediately to the left (lower number of bands) are labeled with an asterisk. For example, phoneme recognition in the 10-band condition was significantly higher than in the 8-band condition for both sets of stimuli. The points beyond which scores were not statistically different from those in the 30-band conditions are indicated by arrows. Increasing the number of bands from 20 to 30 for vowels or

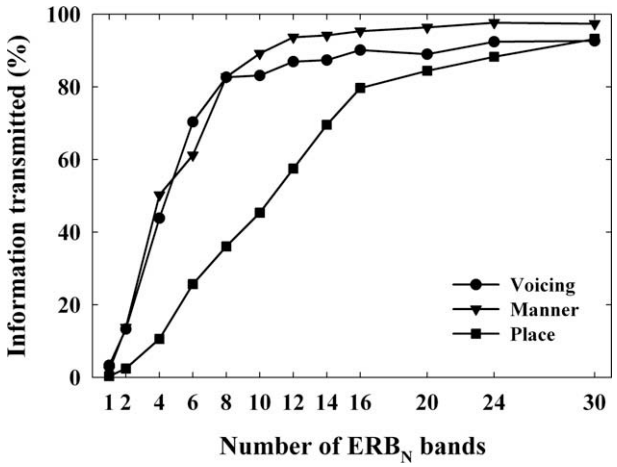

Fig. 3. Percent information transmitted for the features voicing, manner and place of articulation as a function of the number of ERB_N bands (consonant set only).

from 16 to 30 for consonants did not lead to significant improvement, suggesting that close to asymptotic performance was reached in these two number-of-bands conditions.

The average consonant confusion matrices were analyzed in terms of information transmission (Miller and Nicely, 1955).[2] The reception of voicing, manner, and place of articulation was evaluated for each number-of-bands condition. The results of this evaluation are shown in Fig. 3. Transmission of voicing and manner features was very similar. Information transmitted for these two features increased rapidly from nearly 0% in the 1-band condition to 82% in the 8-band condition. Nearly all information concerning these two features was received with only 12 bands. In contrast, reception of place of articulation was only 57% in the 12-band condition. Between 16 and 30 bands, the percentage of information transmitted increased more gradually. Taken together, these results suggest that when 8 speech bands or more were available the inability to reach asymptotic performance may be attributed to a poor transmission of the place feature.

According to the present data, at least half of the 1-ERB_N width bands between 80 and 7563 Hz must convey undistorted speech energy for NH listeners to achieve high levels of phoneme recognition. This contrasts with the results from studies assessing the number spectral channels needed to understand speech using vocoder processing. Vocoder processing is a technique originally developed for speech transmission (Dudley, 1939). It is now often used to simulate cochlear implant (CI) speech processing in NH listeners. Vocoder processing involves dividing an incoming sound into several frequency bands. Then, the slowly-varying amplitude fluctuations (i.e., the temporal envelopes) are extracted from each band and used to modulate a carrier (either noise or tone) having a frequency matching that of the original band. The amount of spectral information is hence determined by the number of frequency bands. Subjects listening to vocoded phonemes usually achieve asymptotic performance with only 4–12 channels of spectral information, suggesting that less than 12 broad channels of spectral information are sufficient to understand speech in quiet (Shannon et al., 1995; Dorman et al., 1997; Loizou et al., 1999; Xu et al., 2005; Apoux and Bacon, 2008a). In contrast, as can be seen in Fig. 2, 16–20 speech bands were necessary to reach maximum performance in this study.

Although this difference may not be surprising given the dissimilarities between the present study and vocoder processing, it is interesting to consider what factors may have contributed to the relatively poor performance observed here. One potential factor is the unpredictable location of the target speech bands in the present study. It is well established that when listeners cannot predict the frequency at which a target tone will be presented, the threshold for detection is higher than when they can. This uncertainty effect, however, is relatively modest (typically 3 dB) and usually disappears if the target tone is preceded by a tonal cue at the same frequency (Gilliom and Mills, 1976; Hübner and Hafter, 1995; Green and McKeown, 2001; Scharf et al., 2007). According to Scharf et al. (2007), the delay between cue onset and target onset must be at least 52 ms to overcome the uncertainty related to the spectral locus of the target tone. In view of that, it may reasonably be argued that the effect of uncertainty, if any, was limited in the present experiment. Indeed, one may consider the initial phoneme in each stimulus as a potential cue for the spectral loci of the target signal, especially for consonant recognition in which the initial vowel duration varied from approximately 150–300 ms. Another potential factor is the exceedingly sparse representation of the speech spectrum produced by the present approach. Indeed, the vocoder data presented in Fig. 2 were all obtained using noise vocoders. Therefore, while fine spectral details were lost, the general shape of the speech spectrum was preserved. It is unlikely, however, that the presence of holes in the spectrum can account for the difference in recognition scores as several studies comparing the intelligibility of speech processed through sine-wave versus noise vocoders failed to show any advantage with the latter, suggesting that the relatively sparse distribution of spectral information does not adversely affect intelligibility (Dorman et al., 1997). The most likely explanation for the difference in performance between the results of the present study and those from vocoder studies is related to the nature of the internal representation resulting from the combined spectral or auditory channels. In vocoder studies, each channel conveys temporal information averaged across a broad frequency region. When combined, the channels encompass the entire speech spectrum so that no frequency region is omitted. In this study, each channel conveyed unprocessed information extracted from a region corresponding to a single auditory filter (as occurs in normal processing). As channels were omitted, the corresponding information was also omitted. Accordingly, vocoder processing may be viewed as providing a comprehensive but blurred representation of the speech spectrum, whereas the present processing may be viewed as providing a partial representation with detailed local information. It seems then from the results of the present study that providing reduced spectral information is less detrimental to intelligibility in quiet than providing partial information.

3.2. Experiment 2: Vowel and consonant recognition in noise

Figs. 4 and 5 display the data for vowels and consonants, respectively. In each figure, the left and right panels correspond to the data for the noise background (SSN) and the speech background (RS) conditions, respectively, with each panel showing percent correct identification score as a function of the number of target speech bands, averaged across listeners for each of the six SNR conditions. For reference, scores obtained in quiet (Experiment 1) as computed for the six subjects tested in that particular condition are indicated by a bold line in each panel.

A separate two-way ANOVA with repeated measures was performed for each speech material/masker combination with factors

[2] An analysis of information transmission was not performed for vowels because reasonable ways to group the stimuli in order to summarize the pattern of confusions could not be found. In particular, duration and formant frequency, the acoustic features commonly considered for vowels, were not consistent across talkers (see Table V in Hillenbrand et al., 1995). For example, several vowels fell in the long duration group when produced by males and in the short duration group when produced by females. Because talker gender was not coded during the experiment, separate analyses for each gender could not be conducted.

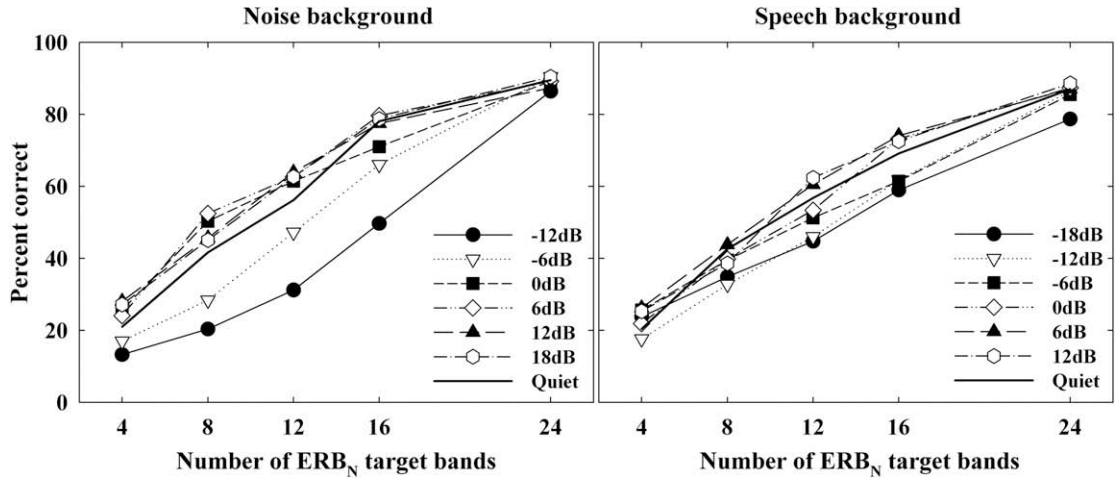

Fig. 4. Percentage of vowels correctly identified as a function of the number of ERB_N target bands. The results obtained in the presence of interleaved speech-shaped noise and time-reversed speech bands are presented in the left and right panels, respectively. The parameter is the SNR.

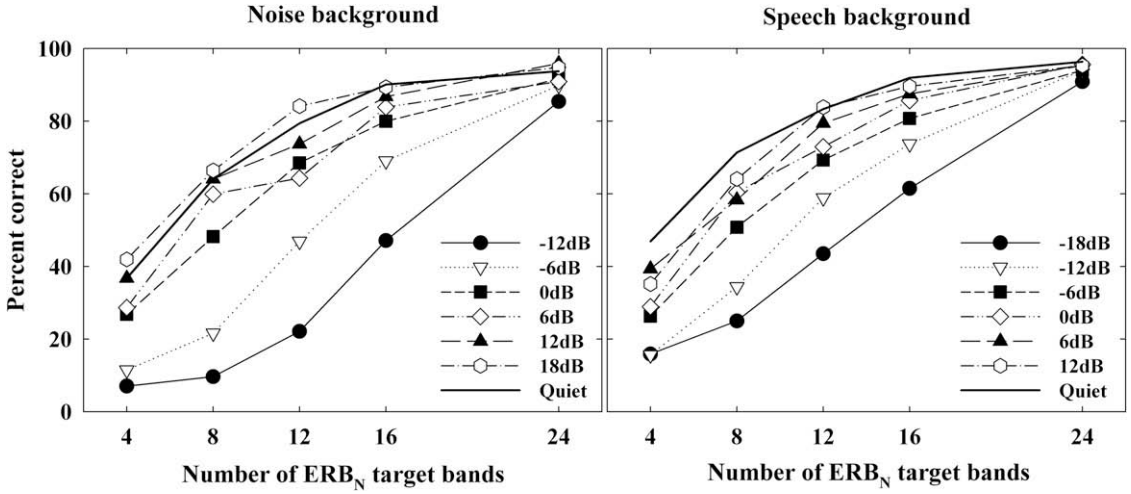

Fig. 5. Same as Fig. 4, except the speech stimuli were consonants.

of number of target bands and SNR. All four analyses indicated a significant effect of number of target bands and a significant effect of SNR (all $p < 0.001$). The interaction between the two main factors was also significant in all four analyses (all $p < 0.001$). Scores tended to be more affected by the presence of the SSN than by that of the RS but overall phoneme identification scores increased as the number of target bands increased (or as the number of masker bands decreased). The effect of the SNR was more complex. Not surprisingly, phoneme recognition with 24 speech bands was little affected by the presence of six interleaved masker bands. With less than 24 speech bands, the three more favorable SNRs did not systematically lead to poorer phoneme recognition scores. A significant drop in performance was systematically observed with the two less favorable SNRs only. This pattern, however, differed slightly as a function of the phoneme set. Vowel recognition was only mildly affected by the presence of interleaved RS (Fig. 4, right panel). Even at −18 dB SNR, the largest drop in performance was only 12 percentage points (12-band condition). In the SSN condition, identification scores remained similar to that measured in quiet except when the SNR was −6 dB or below (Fig. 4, left panel). Although limited, some *improvement* in scores was observed when the interleaved masker bands were added to the vowel stimuli (less than 6 percentage points). This effect presumably reflects spectral restoration (Warren et al., 1997; Apoux and Bacon, 2008b). Consonant recognition was more sensitive to the presence of the masker (Fig. 5). Indeed, identification scores worsened even when the masker bands were added at positive SNRs. However, negative SNRs were still necessary to produce a strong deterioration in performance. For example, identification scores at 0 dB SNR were no more than 12 percentage points below those in quiet for 8 of the 10 number-of-bands conditions. The largest drop in performance was observed in the 12-band condition with the SSN masker at −12 dB SNR (58 percentage points).

The average consonant confusion matrices were analyzed in terms of information transmission. Figs. 6 and 7 show the results of these analyses for the SSN and the RS conditions, respectively. In each figure, the percentage of information transmitted for the features of voicing, manner, and place of articulation is displayed in the top, middle, and bottom panels, respectively. The patterns of results generally follow those observed in quiet (Fig. 3). For SSN, the SNR affected the transmission of voicing and place much in the same way as it affected overall performance. However, manner of articulation appeared *relatively* more susceptible to the influence of interleaved speech-shaped noise. For RS as well, it appears that manner of articulation was relatively susceptible to the presence of an off-frequency masker. The finding that manner of

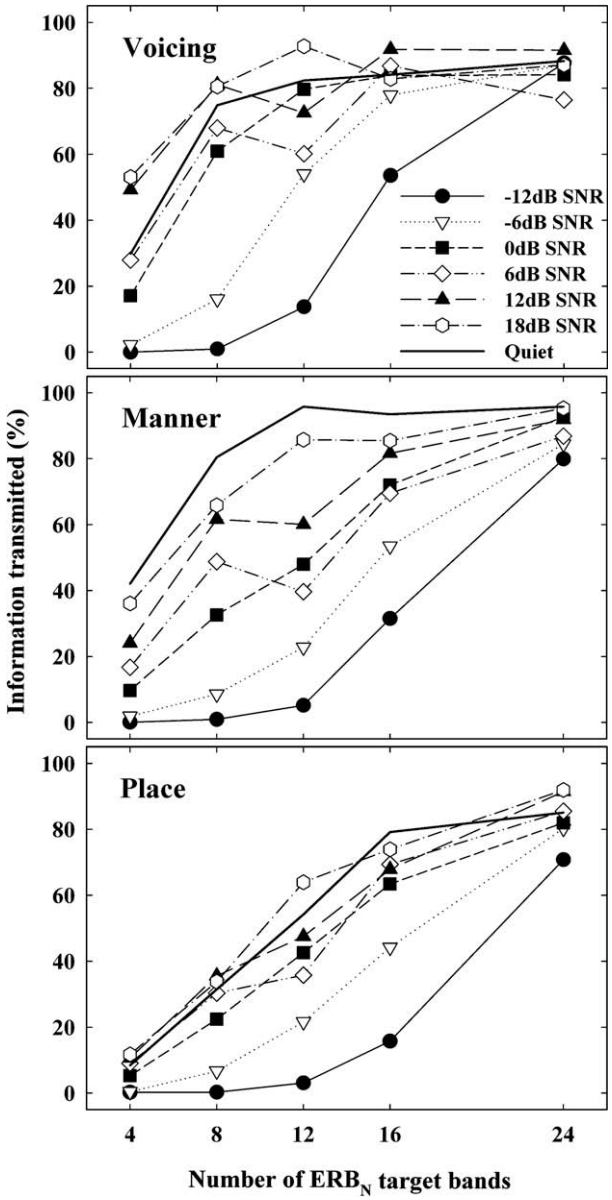

Fig. 6. Percent information transmitted as a function of the number of ERB_N bands in the presence of interleaved speech-shaped noise. The top, middle, and bottom panels show the transmission of voicing, manner and place of articulation, respectively. In each panel, the parameter is the SNR.

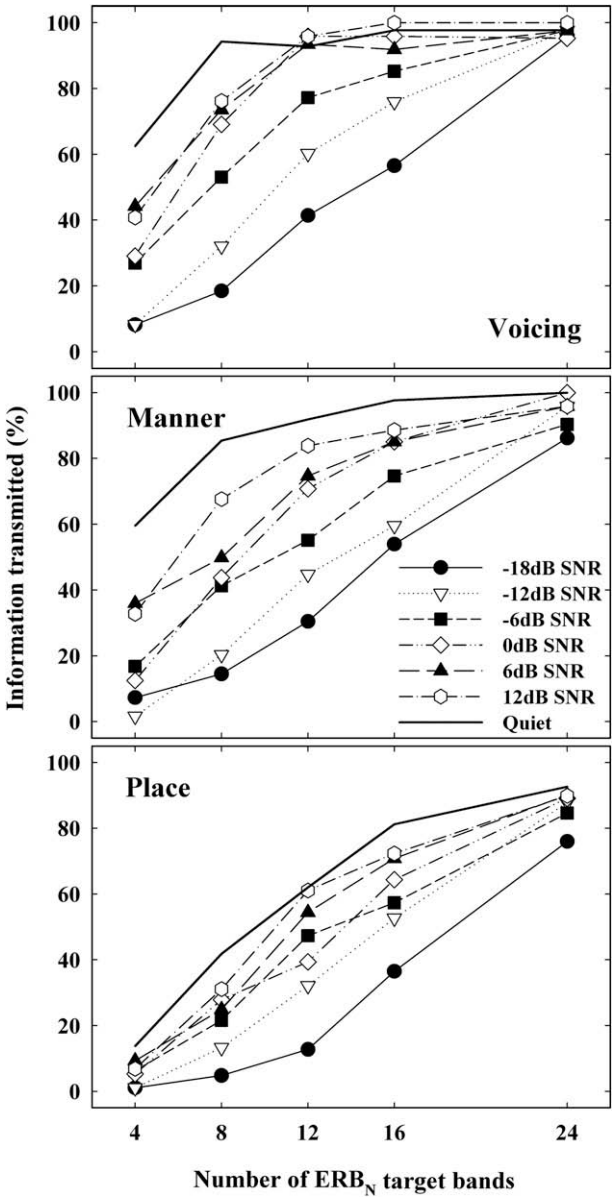

Fig. 7. Same as Fig. 6, except the masker was time-reversed speech.

articulation is more susceptible than voicing to the presence of off-frequency maskers is consistent with the results obtained with on-frequency maskers (e.g., Benkí, 2003).

Fig. 8 presents the "release from interference", the difference between phoneme recognition in fluctuating (RS) versus steady (SSN) off-frequency backgrounds, computed for each combination of number of target bands and SNR. Positive values indicate higher recognition scores in the presence of the fluctuating background (RS). One notable feature of these data is that different patterns were observed for vowels and for consonants. Release from interference was generally limited with vowel stimuli. Also, at positive SNRs, vowel performance was often better in the presence of SSN (negative release values). This latter result may be attributed to spectral restoration as mentioned earlier and suggests that steady state noise is a more efficient spectral "filler" than time-reversed speech. Release from interference was larger with consonants. It generally decreased as the number of speech and masker bands diverged and also with increasing SNR. The largest effect was observed in the 12-band condition at −12 dB SNR (37%).

Taken together, the larger masking effect and release from interference observed with consonants compared to vowels suggests that energy from the masker bands "spilled out" into the speech bands. It is a well established fact that consonant level is considerably lower than vowel level (e.g., Table I in Kennedy et al., 1998). Assuming that noise was present in the speech bands, it is not surprising then that smaller amounts of noise were necessary to interfere with consonant recognition. Indeed, the *effective* SNR in the speech bands was presumably lower (i.e., less favorable) for target consonants than for target vowels. It should be noted, however, that (i) vowel recognition was little affected by the addition of interleaved masker bands and (ii) a fairly small amount of masking release was observed with CVC stimuli, suggesting that the amount of noise present in the speech bands was somewhat limited. The small amount of noise in the speech bands, however, seemed sufficient to observe some amount of masking release with the consonant stimuli.

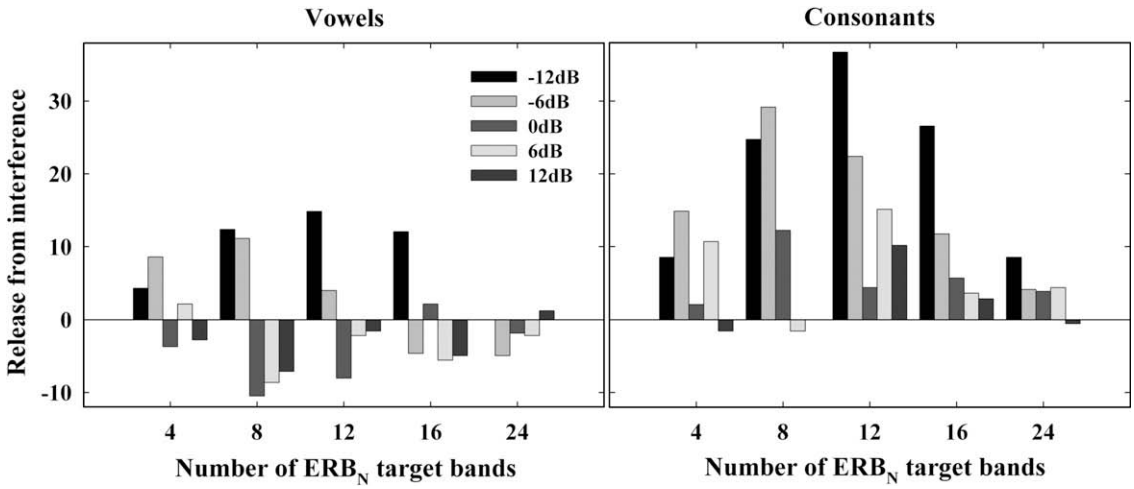

Fig. 8. Release from interference, the difference between phoneme recognition in fluctuating (RS) and steady (SSN) off-frequency backgrounds, as a function of the number of ERB_N target bands, computed for each SNR. Data for vowels and consonants are presented in the left and right panels, respectively. Positive values indicate higher recognition scores in the presence of the fluctuating background (RS).

One motivation for adding noise to the non-speech bands in Experiment 2 was to prevent subjects from using speech information potentially available in these bands. Because the addition of noise led to poorer performance in several conditions, one could argue that subjects were extracting speech cues that spread to these non-speech regions, and therefore the quiet data from Experiment 1 do not reflect the actual number of 1-ERB_N-wide bands needed to achieve the observed performance. A closer look at the data, however, suggests an explanation in terms of on-frequency or "within-speech band" masking. First, one would assume performance to be especially high if subjects had access to more than the prescribed number of bands. Clearly, the results are not consistent with such an assumption as the performance of the subjects was well below that of subjects listening to vocoded speech (see Fig. 1). Second, because speech in the non-speech bands was highly attenuated by the analysis filters, one would expect the addition of even low-level noise to impair performance if subjects were using this information. Therefore, had masking of off-frequency speech occurred, the effect of noise should have been observed at far more positive SNRs. Finally, the release from interference data also support an interpretation in terms of within-speech band masking. In particular, the fact that no masking release was observed for vowels at positive SNRs suggests that subjects were not using speech information in the non-target bands. In view of the above, it is reasonable to conclude that off-frequency speech information did not contribute significantly to overall performance. Accordingly, the results of Experiment 1 should accurately reflect the relationship between speech intelligibility and number of available 1-ERB_N-wide speech bands.

4. Discussion

In Experiment 1, subjects were asked to identify phonemes when presented with a subset of auditory filter width speech bands. In Experiment 2, non-overlapping noise bands were presented simultaneously with the speech bands. Surprisingly, subjects performed similarly in both experiments so long as the noise level remained equal to or lower than that of the target. This finding illustrates the remarkable capacity of the auditory system to effectively process the output of a few auditory filters while ignoring the content of the adjacent filters, demonstrating that auditory filter outputs are processed rather independently along the auditory pathway. To our knowledge, this is one of few studies demonstrating directly what has been called "auditory channel independence" using speech stimuli.

Channel independence as illustrated in the present study has several important implications for our understanding of the mechanisms underlying speech recognition in noise. First, it provides evidence that a strategy used by the auditory system to process speech in noise may consist of the selection of a limited number of auditory filter outputs containing relatively undistorted speech and the combination of these outputs to reconstruct the internal representation of the target signal. Indeed, such a strategy would not be viable without the capacity to process auditory filter outputs independently. As a corollary, the present study supports the glimpsing model of speech recognition in noise discussed in Section 1.

Second, the remarkable channel independence observed in this study can be viewed as an indirect examination of the frequency extent of the regions manipulated by the auditory system to extract speech from noise. In both experiments, the ERB_N scale was used to estimate the auditory filter bandwidths. This scale has been used in many studies and has proven to well characterize listeners' ability to resolve the frequency components of simple sounds. ERB_N values, however, have been derived from psychophysical data obtained with simple stimuli. While there is little reason to believe that the frequency resolution used by the auditory system might differ when processing complex stimuli such as speech, it is worth pointing out that the present data support the relevance of the ERB_N scale for studying speech processing. Indeed, one may reasonably argue that channel independence could only be observed because the ERB_N scale accurately estimates the frequency extent of the auditory filters. In other words, had the auditory system operated at a significantly poorer frequency resolution, the effect of noise should have been observed at greater SNRs because broader auditory filters would have been stimulated by both speech and masker bands. It may then be assumed that ERB_N values provide a fairly accurate estimation of the frequency resolution used by the auditory system when processing speech in complex backgrounds.

A main objective of the present study was to measure the number of auditory filter width speech bands needed to understand speech. A fundamental question raised in the Introduction, however, involved the number of auditory filter outputs needed to understand speech. It was suggested earlier that the comparison between intelligibility in quiet and in interleaved noise may provide some insight regarding the contribution to overall intelligibil-

ity of auditory filters not centered on the speech band. The rational was that if these channels contributed to overall intelligibility, the introduction of noise in these channels resulting from the addition of the noise in the non-speech bands would affect performance. As the presence of interleaved noise did not substantially impair performance, it may be assumed that recognition was primarily based on the output of the auditory filters centered on the speech bands, and the contribution of auditory filter outputs other than those centered on the speech bands was limited. If so, then it may be concluded that the present study reflects the number of auditory filter outputs needed to understand speech.

Although our findings are consistent with current models of speech recognition in noise, one may question the effectiveness of a strategy based on the perception of samples from time–frequency regions containing relatively undistorted speech when as many as 20 1-ERB_N width bands are necessary to correctly identify speech. Moreover, it is likely that a majority of auditory filters would pass a significant amount of noise and might therefore be ignored in real-world situations, resulting in poor speech intelligibility. In everyday settings, however, the auditory system might be able to reconstruct an interpretable representation of a target speech signal with the output of even fewer auditory filters than measured here. Indeed, it has been demonstrated that in adverse conditions the intelligibility of isolated phonemes is poorer than what is typically observed for sentences (Bosman, 1989, reported in Bronkhorst et al., 1993). High level linguistic information (syntactic and semantic), not available when identifying isolated phonemes, plays an important role in the unmasking of speech. Accordingly, it may be assumed that accurate recognition of connected speech necessitates a smaller number of bands than recognition of phonemes. Moreover, it should be noted that significant quantities of information were still transmitted with low numbers of bands in the present experiments. For example, 12 bands were sufficient to identify more than 50% of the vowels and only 6 bands were needed to identify more than 50% of the consonants. Finally, several studies mentioned in the Introduction have demonstrated that listeners are able to extract usable information from frequency regions containing some noise and that even elements of speech below the noise level (i.e., negative SNRs) may still contribute to overall intelligibility. Taken together, the above considerations suggest that a small number of 1-ERB_N width bands and therefore a small number of auditory filter outputs may be sufficient to maintain a communication. The extraction and combination of a limited number of auditory filter outputs containing usable acoustic speech cues therefore remains a plausible strategy for speech recognition in noise.

Finally, the present study has implications for our understanding of the mechanisms underlying poorer-than-normal speech recognition in noise in listeners with sensorineural hearing loss. Damage to outer and inner hair cells is one of the factors frequently evoked to account for this poor performance. There are at least two ways by which damaged hair cells may affect speech recognition in noise. The first way is by smoothing the internal representation of the spectrum. When background noise is present, the smoothing effect is exacerbated because the noise reduces the prominence of spectral peaks, resulting in far greater difficulties. This interpretation is principally related to the fact that listeners with sensorineural hearing loss often have broader auditory filters (e.g., Zwicker and Schorn, 1978). The second way, illustrated in the present study, is related to the reduced number of discrete auditory filters. It is reasonable to assume that because of their broader auditory filters, listeners with moderate to severe hearing loss rely on fewer independent channels. The number of available auditory channels may further be limited by the absence of functional inner hair cells in entire regions of the cochlea. Cochlear implant listeners also rely on a small number of channels of spectral information. This limitation is usually explained in terms of the limited number of physical electrodes and by electrode interactions (CI listeners often cannot distinguish between all the available electrodes because adjacent electrodes stimulate the same population of neurons). Since it is assumed that speech recognition in noise relies on the extraction of a limited number of frequency regions, it is apparent how having access to a limited number of independent auditory channels may affect speech recognition in noise. Indeed, the inability to partition the incoming signal into a large number of independent bands should decrease the probability of uncovering regions in which the target signal is least affected by the background.

The combination of the two above factors provides a credible explanation for the mechanisms underlying the apparent inability of hearing-impaired (HI) and CI listeners to take advantage of momentary, frequency-specific improvements in SNR. Firstly, listeners with reduced frequency selectivity may have access to a limited number of auditory filter outputs. Secondly, most outputs, because they are derived from broader-than-normal auditory filters, pass more noise, further diminishing the probability for a given band to convey undistorted speech. As a consequence, the probability to uncover regions in which the target signal is least affected by the background is greatly diminished in these individuals.

5. Summary

The present study evaluated the overall number of 1-ERB_N width bands necessary to reconstruct an interpretable representation of a target speech signal irrespective of the frequency location of the bands. Twenty such bands were required to accurately identify vowels while "only" 16 speech bands were necessary to identify consonants. Phoneme recognition remained essentially unchanged when bands of speech-shaped noise or time-reversed speech were added in the non-speech bands at positive SNRs. Although limited for vowels, a drop in performance was observed when these masker bands were present at relatively high levels. The following conclusions can be drawn from this study:

i. The capacity of the normal auditory system to effectively process the output of a few auditory filters while ignoring the content of the remaining filters, i.e., channel independence, is remarkably high even when processing complex stimuli such as speech.
ii. The high level of channel independence observed in the present study suggests that the frequency resolution used by the normal auditory system to extract speech from spectrally-adjacent noise is reasonably well estimated by the ERB_N scale.
iii. The high level of channel independence observed in the present study strongly supports the view that a strategy used by the auditory system to extract speech from noise may be to select a limited number of auditory filter outputs containing relatively undistorted speech and to combine these outputs to reconstruct a representation of the target speech signal.
iv. Listeners with moderate to severe hearing loss and CI users, because they presumably rely on fewer independent channels than NH listeners, may experience a significant decrease in the ability to uncover frequency regions in which the target signal is least affected by the background. In other words, a factor limiting the intelligibility of speech in the presence of background noise in HI and CI listeners may be the reduced number of independent auditory or spectral channels typically associated with these listeners. The reduced number of channels should be particularly etrimental in the presence of spectrally-fluctuating backgrounds.

Acknowledgments

The authors thank Lendra Friesen, Michael Dorman and Li Xu for sharing their data. 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.).

References

ANSI S3.5-1969, 1969. Methods for Calculation of the Articulation Index. American National Standards Institute, New York.

ANSI S3.5-1997, 1997. Methods for Calculation of the Speech Intelligibility Index. American National Standards Institute, New York.

ANSI S3.6-2004, 2004. Specifications for Audiometers. American National Standards Institute, New York.

Apoux, F., Bacon, S.P., 2008a. Differential contribution of envelope fluctuations across frequency to consonant identification in quiet. J. Acoust. Soc. Am. 123 (5), 2792–2800.

Apoux, F., Bacon, S.P., 2008b. Selectivity of modulation interference for consonant identification in normal-hearing listeners. J. Acoust. Soc. Am. 123 (3), 1665–1672.

Baskent, D., Shannon, R.V., 2006. Frequency transposition around dead regions simulated with a noiseband vocoder. J. Acoust. Soc. Am. 119 (2), 1156–1163.

Benkí, J.J., 2003. Analysis of English nonsense syllable recognition in noise. Phonetica 60, 129–157.

Bosman, A.J., 1989. Speech perception by the hearing impaired. Doctoral thesis. University of Utrecht, Utrecht, The Netherlands.

Breeuwer, M., Plomp, R., 1984. Speechreading supplemented with frequency-selective sound-pressure information. J. Acoust. Soc. Am. 76 (3), 686–691.

Breeuwer, M., Plomp, R., 1985. Speechreading supplemented with formant-frequency information from voiced speech. J. Acoust. Soc. Am. 77 (1), 314–317.

Breeuwer, M., Plomp, R., 1986. Speechreading supplemented with auditorily presented speech parameters. J. Acoust. Soc. Am. 79 (2), 481–499.

Bronkhorst, A.W., Bosman, A.J., Smoorenburg, G.F., 1993. A model for context effects in speech recognition. J. Acoust. Soc. Am. 93 (1), 499–509.

Brungart, D.S., Chang, P., Simpson, B., Wang, D., 2006. Isolating the energetic component of speech-on-speech masking with ideal time–frequency segregation. J. Acoust. Soc. Am. 120 (6), 4007–4018.

Buss, E., Hall, J.W., Grose, J.H., 2003. Spectral integration of synchronous and asynchronous cues to consonant identification. J. Acoust. Soc. Am. 115 (5), 2278–2285.

Celmer, R.D., Bienvenue, G.R., 1987. Critical bands in the perception of speech signals by normal and sensorineural hearing loss listeners. In: Schouten, M.E.H. (Ed.), The Psychophysics of Speech Perception. Nijhoff, Dordrecht, pp. 473–480.

Cooke, M.P., 2003. Glimpsing speech. J. Phonetics 31 (57), 9–584.

Cooke, M.P., 2005. Making sense of everyday speech: a glimpsing account. In: Divenyi, P. (Ed.), Speech Separation by Humans and Machines. Kluwer Academic, Dordrecht, pp. 305–314.

Cooke, M.P., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119 (3), 1562–1573.

Dorman, M.F., Loizou, P.C., Rainey, D., 1997. Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. J. Acoust. Soc. Am. 102 (4), 2403–2411.

Drullman, R., 1995. Speech intelligibility in noise: relative contribution of speech elements above and below the noise level. J. Acoust. Soc. Am. 98 (3), 1796–1798.

Dubno, J.R., Horwitz, A.R., Ahlstrom, J.B., 2002. Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing. J. Acoust. Soc. Am. 111 (6), 2897–2907.

Dudley, H., 1939. Remaking speech. J. Acoust. Soc. Am. 11 (2), 169–177.

Festen, J.M., Plomp, R., 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J. Acoust. Soc. Am. 88 (4), 1725–1736.

Fletcher, H., 1940. Auditory patterns. Rev. Mod. Phys. 12, 47–65.

French, N.R., Steinberg, J.C., 1947. Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Am. 19 (1), 90–119.

Friesen, L.M., Shannon, R.V., Baskent, D., Wang, X., 2001. Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. J. Acoust. Soc. Am. 110 (2), 1150–1163.

Füllgrabe, C., Berthommier, F., Lorenzi, C., 2006. Masking release for consonant features in temporally fluctuating background noise. Hear. Res. 211 (1–2), 74–84.

Gilliom, J.D., Mills, W.M., 1976. Information extraction from contralateral cues in the detection of signals of uncertain frequency. J. Acoust. Soc. Am. 59 (6), 1428–1433.

Glasberg, B.R., Moore, B.C.J., 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 47 (1–2), 103–138.

Grant, K.W., Braida, L.D., 1991. Evaluating the articulation index for audiovisual input. J. Acoust. Soc. Am. 89 (6), 2952–2960.

Green, T.J., McKeown, J.D., 2001. Capture of attention in selective frequency listening. J. Exp. Psychol. Hum. Percept. Perform. 27 (5), 1197–1210.

Gustafsson, H.A., Arlinger, S.D., 1994. Masking of speech by amplitude modulated noise. J. Acoust. Soc. Am. 95 (1), 518–529.

Healy, E.W., 1998. A minimum spectral contrast rule for speech recognition: intelligibility based upon contrasting pairs of narrow-band amplitude patterns. Doctoral dissertation. University of Wisconsin-Milwaukee, Milwaukee, WI, USA.

Healy, E.W., Warren, R.M., 2003. The role of contrasting temporal amplitude patterns in the perception of speech. J. Acoust. Soc. Am. 113 (3), 1676–1688.

Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. 97 (5), 3099–3111.

Howard-Jones, P.A., Rosen, S., 1993. Uncomodulated glimpsing in 'checkerboard' noise. J. Acoust. Soc. Am. 93 (5), 2915–2922.

Hübner, R., Hafter, E.R., 1995. Cuing mechanisms in auditory signal detection. Percept. Psychophys. 57 (2), 197–202.

Iyer, N., Brungart, D.S., Simpson, B.D., 2007. Effects of periodic masker interruption on the intelligibility of interrupted speech. J. Acoust. Soc. Am. 122 (3), 1693–1701.

Kalikow, D.N., Stevens, K.N., 1977. Development of a test of speech intelligibility in noise using sentence material with controlled word predictability. J. Acoust. Soc. Am. 61 (5), 1337–1351.

Kennedy, E., Levitt, H., Neuman, A.C., Weiss, M., 1998. Consonant–vowel intensity ratios for maximizing consonant recognition by hearing-impaired listeners. J. Acoust. Soc. Am. 103 (2), 1098–1114.

Kryter, K.D., 1962. Validation of the articulation index. J. Acoust. Soc. Am. 34 (11), 1698–1702.

Li, N., Loizou, P., 2008. Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction. J. Acoust. Soc. Am. 123 (3), 1673–1682.

Lippman, R.P., 1996. Accurate consonant perception without mid-frequency speech energy. IEEE Trans. Speech Audio Process. 4 (1), 66–69.

Loizou, P.C., Dorman, M., Tu, Z., 1999. On the number of channels needed to understand speech. J. Acoust. Soc. Am. 106 (4), 2097–2103.

Miller, G.A., Licklider, J.C.R., 1950. The intelligibility of interrupted speech. J. Acoust. Soc. Am. 22 (2), 167–173.

Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27 (2), 338–352.

Nelson, P.B., Jin, S.-H., Carney, A.E., Nelson, D.A., 2003. Understanding speech in modulated interference: cochlear implant users and normal-hearing listeners. J. Acoust. Soc. Am. 113 (2), 961–968.

Rhebergen, K.S., Versfeld, N.J., Dreschler, W.A., 2005. Release from informational masking by time reversal of native and non-native interfering speech (L). J. Acoust. Soc. Am. 118 (3), 1274–1277.

Shannon, R.V., Jensvold, A., Padilla, M., Robert, M.E., Wang, X., 1999. Consonant recordings for speech testing. J. Acoust. Soc. Am. 106 (6), L71–74.

Shannon, R.V., Zeng, F., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270 (5234), 303–304.

Scharf, B., Reeves, A., Suciu, J., 2007. The time required to focus on a cued signal frequency. J. Acoust. Soc. Am. 121 (4), 2149–2157.

Warren, R.M., Bashford Jr., J.A., Lenz, P.W., 2004. Intelligibility of bandpass filtered speech: steepness of slopes required to eliminate transition band contributions. J. Acoust. Soc. Am. 115 (3), 1292–1295.

Warren, R.M., Hainsworth, K.R., Brubaker, B.S., Bashford, J.A., Healy, E.W., 1997. Spectral restoration of speech: intelligibility is increased by inserting noise in spectral gaps. Percept. Psychophys. 59 (2), 275–283.

Warren, R.M., Riener, K.R., Bashford, J.A., Brubaker, B.S., 1995. Spectral redundancy: intelligibility of sentences heard through narrow spectral slits. Percept. Psychophys. 57 (2), 175–182.

Xu, L., Thompson, C.S., Pfingst, B.E., 2005. Relative contributions of spectral and temporal cues for phoneme recognition. J. Acoust. Soc. Am. 117 (5), 3255–3267.

Zwicker, E., Schorn, K., 1978. Psychoacoustical tuning curves in audiology. Audiology 17 (4), 120–140.