Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Phoneme recognition as a function of the number of auditory filter outputs

Frederic Apoux and Eric Healy

University of South Carolina, Speech Psychoacoustics Laboratory, Department of Communication Sciences and Disorders, William Brice Bldg., 1621 Greene St., Columbia, SC 29208, USA
apoux@sc.edu

It has been proposed that listeners take advantage of brief "coups d'oeil" when processing speech in noise. These glimpses can be characterized both in time and frequency. The obligatory role of the auditory filters in determining the nature of any further processing suggests that the frequency extent of a glimpse should be equivalent to that of an auditory filter. Accordingly, it is hypothesized that the spectral characteristics of glimpses primarily relate to the available number of auditory channels. The present study investigated the number of auditory filter outputs needed to identify phonemes in quiet. Stimuli were first restricted to 80-7563 Hz and then split into 30 contiguous auditory filter width bands. Normal-hearing listeners were presented with *N* bands having spectral locations selected randomly from trial to trial. No signal was presented in the other bands. Consistent with previous studies, performance gradually increased as the number of bands increased. An asymptote was reached with 24 and 16 bands for vowels and consonants, respectively. While high levels of speech understanding are typically observed with as few as 4 channels of spectral information, our results indicate that accurate phoneme recognition requires combination of a much larger number of auditory filter outputs.

1 Introduction

A fundamental property of the peripheral auditory system is that it operates as a kind of frequency analyzer [1]. Because of their obligatory role in determining the nature of any further processing, much effort has gone into determining the role of these "auditory filters." One evident role is to resolve spectral peaks in the speech signal. This role is consistent with the traditional Fourier-based approach to speech understanding, in which speech information is conveyed by the distribution of energy along the audio-frequency axis. Indeed, vowel and consonant identity is specified to some extent by the location of spectral peaks (i.e., the formants) and small differences in formant frequencies can lead to changes in phonetic identities. The auditory system would not be able to analyze spectral shapes and changes in those spectral shapes across time without the capacity to resolve spectral features in the acoustic signal.

A number of studies have investigated the effects of reducing spectral resolution on speech recognition in quiet in normal-hearing (NH) listeners [2,3,4]. Data typically show reduced speech intelligibility in conditions of severely degraded frequency selectivity. However, speech intelligibility is not substantially affected by mild-to-moderate spectral smearing, suggesting that only *broadly tuned* auditory filters are needed to understand speech in quiet. Consistent with the finding that fine spectral resolution is not critical to understand speech in quiet, NH listeners hearing simulations of cochlear implant (CI) signal processing (i.e., vocoder) usually achieve asymptotic performance with 4 to 12 channels of spectral information [5,6,7]. In other words, 12 broad auditory filters are sufficient to understand speech in quiet. Although vocoder and spectral smearing studies do not exclude a role of frequency selectivity in speech recognition, they demonstrate that *the ear's ability to resolve spectral contrasts is much larger than would be required to understand speech in quiet.*

Since the ear's *high* spectral resolution does not play a critical role in speech recognition in quiet, it has been suggested that it may be of particular importance to understanding speech in noisy environments. The notion that fine spectral resolution plays a role in the unmasking of speech is consistent with current views of speech recognition in noise. One view suggests that the strategy used by NH listeners to understand speech in noise is to listen in spectral and temporal dips [8]. The "listening-in-the-dips" hypothesis is supported by the results of many behavioral studies showing that intelligibility in noise increases substantially when spectral and/or temporal gaps are introduced in the masker [9]. More recently, Cooke [10] suggested that speech recognition in noise does not rely solely on momentary improvements in signal-to-noise ratio (SNR) but more generally on the ability to extract and combine information from spectro-temporal regions that contain a reasonably undistorted view of local signal properties, the so-called glimpses.

The views discussed above are based on two findings. First, the success of NH listeners in recognizing speech in noise can be attributed almost exclusively to the relationship between speech and noise intensities, i.e., the SNR. Because most natural sounds, such as speech, are highly modulated both in time and frequency, this relationship is usually non-uniform across frequency and may change rapidly over brief periods of time. As a result, there is always a fair probability to observe frequency regions dominated by the target speech at any moment in time. Moreover, speech information is highly redundant in the frequency domain, in that individual speech cues can be degraded to various degrees without affecting overall speech recognition. One consequence of speech redundancy is that the combination of even a small number of frequency regions dominated by the target speech may be sufficient to maintain a communication. Some degree of spectral resolution is needed, however, to take advantage of undistorted views of local signal properties.

The above considerations have led to the idea that a primary function of the ear's *fine* spectral resolution is to support the segregation of simultaneous acoustic signals. In this view, the peripheral auditory system has a passive but critical role. This role, consistent with the models of speech recognition in noise discussed earlier, is to partition the incoming sound mixture into a series of bands so that spectro-temporal regions dominated by the target speech may be uncovered and grouped together at a latter processing stage to form the internal representation of the signal of interest. The importance of fine spectral resolution is therefore apparent. Indeed, *the possibility to partition the incoming signal into a large number of bands should increase the probability of uncovering frequency regions in which the target signal is least affected by the background.* Another consequence is that the auditory system would reconstruct a representation of the target signal by combining the output of auditory filters dominated by the target speech. Accordingly, the goal of the present study was to investigate the relationship between speech intelligibility and number of available auditory filter outputs (or auditory channels) to determine the amount of information needed to reconstruct a representation of a target speech signal. Indeed, the amount of information available to a listener should correspond to the number of auditory filter outputs containing "undistorted" speech, and therefore may be determined by measuring speech recognition as a function of the number of "clean" auditory filter outputs. In the present study, subjects were presented with N speech bands selected randomly from trial to trial among 30 possible auditory filter width bands [11]. In an effort to eliminate the influence of noise in the target speech channels, no noise was added. However, it was anticipated that subjects might be able to use off-frequency information from adjacent non-speech bands. Therefore, to limit the contribution of frequencies lying outside the nominal bandwidth, a second experiment was designed in which noise was presented simultaneously with the target speech. The target and masker bands were interleaved so that overlap in the spectral domain (i.e., peripheral masking) was limited and speech would remain as undistorted as possible. The interleaving arrangement is also a more realistic condition in that listeners are forced to select the output of a limited number of channels and ignore the others.

2 Method

2.1 Subjects

Twelve NH subjects participated. Their ages ranged from 22 to 26 years. Normal hearing was defined as having pure-tone air-conduction thresholds of 20 dB HL or better for octave frequencies from 125 to 8000 Hz. All participants were native speakers of American English and received course credit for their participation.

2.2 Speech material and processing

The stimuli consisted of 9 vowels in /h/-vowel-/d/ environment produced by six speakers (three for each gender) for a total of 54 consonant-vowel-consonants (CVCs), and 16 consonants in /a/-consonant-/a/ environment produced by four speakers (two for each gender) for a total of 64 vowel-consonant-vowels (VCVs). The background noise was a simplified speech spectrum-shaped noise (constant spectrum level below 800 Hz and 6 dB/oct roll-off above 800 Hz). The duration of the masker was always equal to target duration. Prior to combination, speech and noise stimuli were filtered into 30 contiguous frequency bands ranging from 80 to 7563 Hz. Each band was one equivalent rectangular bandwidth ($\mathrm{ERB_N}$) wide so that the filtering simulated the frequency selectivity of the normal auditory system. In the noise conditions, target and masker bands were interleaved. The target was normalized and calibrated so that its overall A-weighted output level was 65 dB when presented alone in the 30 band condition. The overall level of the 30 combined masker bands was adjusted to achieve a specific SNR when compared to the 30 target bands. The value referred to as SNR in the rest of the study corresponds to the SNR computed for the broadband signals. As a consequence, the actual overall level and SNR most likely differed from their initial value in most experimental conditions.

2.3 Procedure

The experimental sessions were organized in the following way. First, phoneme recognition was measured in quiet. Twelve number-of-bands conditions were tested. In each condition, subjects were presented with N $\mathrm{ERB_N}$ speech bands (N = 1, 2, 4, 6, 8, 10, 12, 14, 16, 20, 24 or 30) having locations selected randomly from trial to trial among the 30 possible bands. As a result, speech information was randomly distributed across frequency. Although the amount of information (number of bands) remained constant within a condition, the distribution of this information varied from trial to trial. No signal was presented in the non-speech bands. Six subjects were asked to identify vowels and six subjects were asked to identify consonants. Each subject completed 12 blocks, with each block corresponding to a (randomly ordered) number-of-bands condition. In each block, all 54 CVCs or 64 VCVs were presented once in random order. Then, phoneme recognition was measured with the interleaved noise. Each subject performed the task with the same set of stimuli used in the quiet experiment (CVCs or VCVs). The SNR ranged from -12 to 18 dB in 6-dB steps. All combinations of five numbers of target bands (N = 4, 8, 12, 16 or 24) and six SNRs were tested. Again, each block contained the 54 or 64 stimuli once each, and speech-band locations were determined randomly from trial-to-trial.

Listeners were tested individually in a single-walled, sound-attenuated booth. Stimuli were presented to the listeners binaurally through Sennheiser HD 250 Linear II circumaural headphones. The experiments were performed using 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.

3 Results and discussion

3.1 Quiet data

The averaged data are presented in Fig. 1. The upper and lower panels show the results for vowels and consonants, respectively. As expected, performance increased with increasing numbers of bands. Maximum performance was reached in the 30-band condition for both speech materials. Identification scores were generally lower for vowels than for consonants and the slope of the function was shallower for vowels. As a consequence, 24 bands were necessary to reach asymptotic performance when identifying vowels while only 16 bands were needed when identifying consonants. Separate one-way analyses of variance (ANOVA) with repeated measures indicated that the identification of vowels and consonants was significantly affected by the number of bands (all $p < 0.001$). Post-hoc tests (Tukey) confirmed that identification of vowels and consonants did not increase significantly from 24 to 30 and from 16 to 30 bands, respectively.

An important feature of the above results is that the number of bands necessary to reach asymptotic performance was much larger than what is typically observed in vocoder studies [5,6,7]. A possible explanation for this discrepancy relates to how spectral information was reduced in each group of studies. In the present experiment, spectral information was reduced by creating holes in the spectrum. In vocoder studies, stimuli were processed so that averaged envelope information computed from a broad frequency region was fed to several contiguous auditory channels. Therefore, information was averaged across frequency in these studies while entire regions of the speech spectrum were absent in the present experiment. This difference may very well account for the poorer performance observed in the present study and suggest that creating holes in the spectrum is more damaging to intelligibility than averaging envelope information across frequency as in vocoder processing.

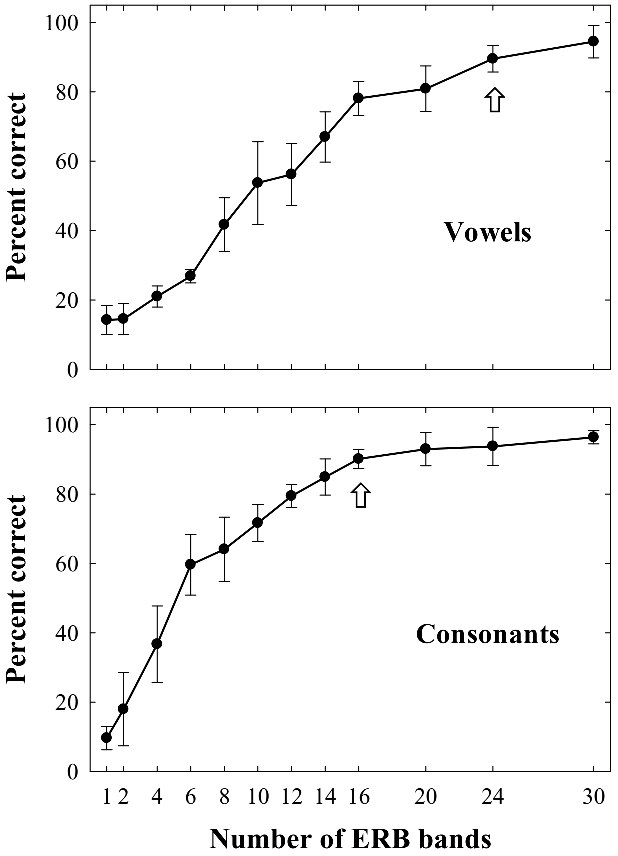

Fig. 1 Percentage of vowels (upper panel) and consonants (lower panel) correctly identified as a function of the number of bands. In each panel, the asymptotic performance is indicated by an arrow. Errors bars indicate one standard deviation.

3.2 Noise data

Figures 2 and 3 show the average percent correct scores as a function of the number of target bands for varying background noise levels for vowels and consonants, respectively. For reference, the averaged identification scores measured in quiet are plotted in each figure. Consistent with the results observed in quiet, intelligibility increased with increasing number of target bands. The effect of the SNR was less clear since the addition of noise in the non-speech bands did not systematically lead to poorer performance. Indeed, comparison between data obtained in quiet and in noise indicates a limited effect of noise at high SNRs. This last result is not uncommon as performance does not systematically drop at very favorable SNRs. However, the SNR at which recognition began to decrease noticeably was quite low in the present study, when compared to what is typically observed. This absence of effect presumably reflects the considerable independence of auditory channels. It is apparent when comparing Figures 2 and 3 that consonant recognition was more affected than vowel recognition by the presence of the off-frequency masker at a given SNR. A closer inspection of the data, however, suggests that the same general masking pattern was observed for both speech materials. In general, phoneme recognition was only mildly affected by the presence of the interleaved masker at SNRs of 0 dB and above. A substantial drop in performance was only observed when the noise was added at very high levels (-6 and -12 dB SNR).

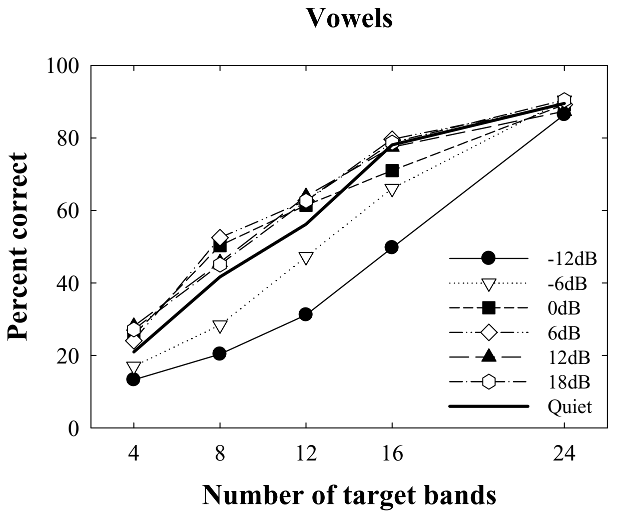

Fig. 2 Percentage of vowels correctly identified as a function of the number of bands. The parameter is the SNR.

Again, a separate two-way ANOVA with repeated measures was performed on each set of data. Both analyses revealed a significant effect of number-of-bands and SNR (all $p < 0.001$). The interaction between the two main factors was also significant (all $p < 0.001$). This significant interaction may be attributed, at least partly, to the limited effect of the 6 noise bands when interleaved with 24 speech bands.

The masking pattern observed in the present experiment strongly suggests an explanation in terms of within-channel masking. At low masker levels, very little noise passed through the speech channels resulting in limited interference. At high masker levels, more noise passed through the speech channels resulting in reduced intelligibility. This explanation is consistent with previous works suggesting a limited influence of off-frequency noise on speech intelligibility [12,13]. It also accounts for the greater sensitivity of the consonant test to the presence of off-frequency maskers. As described previously, the overall level of the masker was adjusted to achieve a specific SNR when compared to the level of the broadband target stimuli as measured across their entire duration (i.e., including both vowels and consonants comprising a syllable). Because consonant level is lower than vowel level [14], overall level was predominantly driven by the vocalic portions. As a consequence, the effective SNR was presumably higher for *target* vowels than for *target* consonants, resulting in a greater effect of noise on consonants than on vowels at "equal" SNRs. The overall SNR, however, was identical for CVCs and VCVs. Considering that most of the disruptive effect of interleaved noise at high SNRs is attributable to within-channel masking, it is reasonable to assume that subjects did not use off-frequency information when performing the task in quiet. Therefore, the quiet data should reflect accurately the relationship between number of auditory filter outputs and phoneme recognition.

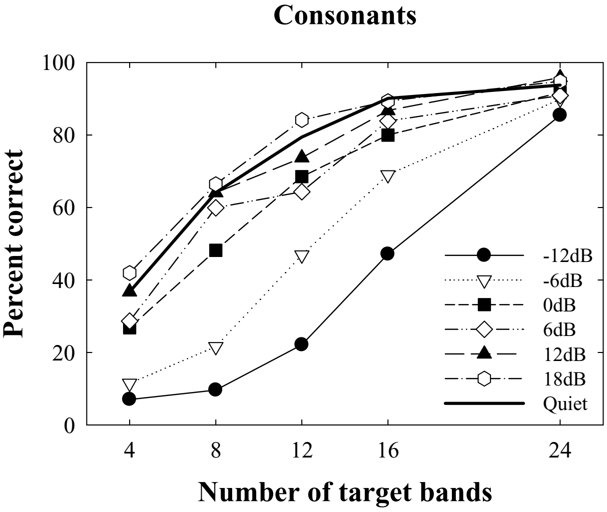

Fig. 3 Percentage of consonants correctly identified as a function of the number of bands. The parameter is the SNR.

Consistent with the power-spectrum model [1,15], the noise data suggest a large independence between auditory channels, as the presence of noise at high levels in some auditory channels does not affect the processing of speech in other channels. These results also validate to some extent the use of the $\mathrm{ERB}_{\mathrm{N}}$ when studying speech processing by the human auditory system. Indeed, the fact that listeners had only small difficulties understanding speech in the presence off-frequency noise suggests that the target and the masker were effectively separated in the auditory system. In other words, the limited peripheral masking suggests that $\mathrm{ERB}_{\mathrm{N}}$ values provided a good estimate of the spectral resolution used by the auditory system when processing speech.

Finally, it is apparent from these results that listeners had very little difficulty combining partial information across frequency. Even in conditions in which the number of auditory channels conveying speech information was very small (i.e., 4 or 8 bands), performance was not affected by the presence of the interleaved off-frequency noise. These findings support our initial assumption that a viable strategy when listening to speech in noise is to select a subset of auditory channels with undistorted speech and group them together to form the internal representation of the target sound. These findings also provide useful information regarding how the ear's *high* spectral resolution may be of particular importance in understanding speech in noisy environments.

4 Conclusions

The present study was designed to evaluate the number of auditory filter outputs containing "undistorted" speech necessary to recognize vowels and consonants. The results indicated that 16 auditory filter width bands are needed to achieve near perfect consonant recognition and that as many as 24 bands are needed to achieve near perfect vowel recognition. The difference in the number of bands needed to identify vowels and consonants may be attributed to the fact that vowel identity is believed to rely primarily on the transmission of spectral cues while consonant identity also relies on the transmission of temporal cues. The number of auditory filter width bands needed to identify phonemes is considerably larger than the number of vocoder channels needed to achieve similar performance. This discrepancy presumably reflects the exclusion of entire frequency regions in the present study.

Phoneme recognition is not substantially affected by the presence of noise in the non-speech bands, irrespective of the available number of bands with undistorted speech. However, interleaved background noise may significantly affect phoneme recognition at high SNRs by spilling over into the speech bands. Taken together, these results suggests that a viable strategy for understanding speech in noise is to select and combine the outputs of a limited number of auditory filters containing undistorted speech, and therefore supports the assumption of a role of peripheral filtering in the unmasking of speech.

Acknowledgments

This research was supported by NIDCD grant No. DC008594.

References

[1] H. Fletcher, "Auditory patterns", *Rev. Mod. Phys.* 12 47-65 (1940).

[2] J. R. Dubno, M. F. Dorman, "Effects of spectral flattening on vowel identification", *J. Acoust. Soc. Am.* 82, 1503-1511 (1987).

[3] T. Baer, B. C. J. Moore, "Effects of spectral smearing on the intelligibility of sentence in noise", *J. Acoust. Soc. Am.* 94, 1229-1241 (1993).

[4] Y. Nejime, B. C. J. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise", *J. Acoust. Soc. Am.* 102, 603-615 (1997).

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

[6] L. Xu, C. S. Thompson, B. E. Pfingst, "Relative contributions of spectral and temporal cues for phoneme recognition", *J. Acoust. Soc. Am.* 117, 3255-3267 (2005).

[7] F. Apoux, S. P. Bacon, "Differential contribution of envelope fluctuations across frequency to consonant identification in quiet", *J. Acoust. Soc. Am.* 123, 2792-2800 (2008a).

[8] G. A. Miller, J. C. R. Licklider, "The intelligibility of interrupted speech", *J. Acoust. Soc. Am.* 22, 167-173 (1950).

[9] C. Fullgrabe, F. Berthommier, C. Lorenzi, "Masking release for consonant features in temporally fluctuating background noise", *Hear. Res.* 211, 74-84 (2006).

[10] M. Cooke, "A glimpsing model of speech in noise", *J. Acoust. Soc. Am.* 119, 1562-1573 (2006).

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

[12] B. J. Kwon, C. W. Turner, "Consonant identification under maskers with sinusoidal modulation: Masking release or modulation interference?", *J. Acoust. Soc. Am.* 110, 1130-1140 (2001).

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

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

[15] R. D. Patterson, B. C. J. Moore, "Auditory filters and excitation patterns as representations of frequency resolution", in *Frequency selectivity in Hearing*, edited by B. C. J. Moore (Academic, London), pp. 123-177 (1986).