Effects of Preceding Noise on the Perception of Voiced Plosives

T. Cervera, W. A. Ainsworth*
Departamento de Psicología Básica, Facultad de Psicología, Universidad de Valencia, Avda. Blasco Ibáñez 21, 46010 Valencia, Spain. Teresa.Cervera@uv.es

Summary

The intelligibility of voiced plosives with added continuous noise is greater than with no noise which is coincident with the speech. In order to investigate this further, experiments have been performed to measure the intelligibility of English plosive-vowel syllables with different durations of preceding noise. The effects of the spectral shape of the noise have also been examined. It was found that, on average, 800 ms of filtered noise preceding the syllable increased the recognition rate of plosives compared with coincident noise. The rate of increase was 14% for band-stop filtered noise, 12% for band-pass, 7% for low-pass, but only 3% for high-pass filtered noise. It is suggested that the mechanism responsible for these increases is the same as that underlying the psychoacoustical ‘enhancement’ effect, a form of adaptation by which new arriving information in one part of the frequency spectrum can receive perceptual salience, relative to the energy at pre-existing frequencies. The effects of various bands of coincident and preceding noise on the recognition of individual plosive-vowel syllables have been examined in conjunction with an acoustic analysis of the syllables. It is concluded that the effects of increasing the duration of the preceding noise are similar to decreasing the level of the noise. These results can provide better understanding about the mechanisms underlying speech perception in noise.

PACS no. 43.71.Gv

1. Introduction

The effects of noise on the perception of speech were studied by Fletcher [1] who showed that in order to maintain the same intelligibility the speech level had to be increased by approximately the same amount as the level of the added noise. Miller, Heise and Lichten [2] showed that intelligibility also depended on the size and type (digits, monosyllables, etc.) of the vocabulary. Confusions amongst consonants were measured by Miller and Nicely [3] who showed that, when noise was added, sounds sharing place of articulation were more likely to be confused than those with a different place of articulation. Consonant confusions in noise have also been studied by Singh and Black [4], Wang and Bilger [5] and Bell, Dirks and Carterette [6]. The noise characteristics and speech presentation levels were found to affect the pattern of consonant confusions.

More recently Ainsworth and Meyer [7] compared the perception of speech in continuous noise with speech in noise which was turned on at the start of the speech and off at the end (coincident noise). They found that for a given

SNR (they employed SNRs of 0 dB and 6 dB) the intelligibility of plosives and vowels in consonant-vowel syllables was some 15% higher in the continuous noise than in the coincident noise. They suggested that this may be due to a change in the firing threshold of neurones in the cochlear nerve brought about by the noise. Evans [8], Costalupes et al. [9] and Gibson et al. [10] reported that as the noise level is increased the threshold of the receptive fields for tones increases and the spike discharge rate decreases. According to Evans these effects do not occur if the noise is presented simultaneously with the test tones for short periods. Ainsworth and Meyer developed an auditory model incorporating these effects which, when used as the front-end of a speech recogniser, gave similar results to human listeners with the same speech stimuli. This physiologically based computational model contained simulations of units in both the cochlear nerve and cochlear nucleus. It was interfaced to an HMM recogniser and tested with the same stimuli that were employed in the human perception experiments.

Delgutte [11] has analysed representations of speech-like sounds in the presence of background noise in the discharge patterns of cochlear nerve fibres. He found that the discharge rate is decreased in the presence of noise, with the decrease being largest just after onset of a tone burst. The addition of background noise results in a reduced prominence of the time course of short-term adaptation in the response pattern. Carney and Geisler [12]
have analysed cochlear nerve fibre responses to spoken plosive-vowel syllables. They found that several acoustic correlates of perceptual features, such as initial spectrum, formant transitions and voice-onset time, are represented in the temporal properties of the fibre responses. The responses of high-spontaneous-rate [13] and low-spontaneous-rate [14] cochlear nerve fibres to plosive-vowel syllables in noise have also been studied.

Although recognition is different from signal detection, both involve some of the same processing. McDaid [15] showed that the masking-level difference (MLD) for a tone burst in a coincident noise burst is smaller than in continuous noise. That is, the signal level required for detecting the presence of the tone in noise of the same duration as the signal is higher than for detecting a tone in continuous noise. He also showed that the masking noise must be activated approximately 500 ms prior to the signal before the maximum MLD occurred. Yost [16], Schlauch, Lanthier and Neve [17], and others, have also investigated duration effects with forward masking of tones by noise.

With speech recognition it might also be expected that the auditory system takes time to adapt to the noise preceding the syllable. Ainsworth [18] attempted to estimate the time constant of this adaptation process by measuring the recognition score as a function of the duration of the noise preceding the speech syllable. The results, however, were inconclusive. Although the recognition rate eventually increased as the duration of the noise was lengthened, it was found that the recognition rate decreased for noise of short duration (of the order of 100 ms). There were, however, a number of differences between these experiments and those of Ainsworth and Meyer [7]. In this earlier experiment the test syllables were plosive-vowel syllables spoken by a French speaker and native French listeners were employed whereas, in the Ainsworth [18] experiment, the syllables were spoken by a British English speaker and English speaking listeners were employed. More importantly perhaps, white noise low-pass filtered with a first-order Butterworth filter with a cut-off point of 500 Hz was used in the first experiment and unfiltered white noise in the second.

In order to investigate the effects of the spectrum shape and duration of preceding noise on the perception of voiced plosives, a new series of experiments has been performed. It is expected that the spectrum of the noise produces different effects on the perception of the plosives. It is known that the frequency region where preceding noise is presented can have important effects on the audibility of a signal. Several Authors [19, 20, 21, 22, 23, 24, 25, 26] showed that an “enhancement effect” is observed when prior exposure to noise is presented in the frequency region around the target frequency. A harmonic complex with one component missing preceding the same harmonic complex with the missing component included (flat spectrum) causes that component to become more prominent. This effect also occurs if the precursor is replaced by band-stop noise with the stop-band corresponding to the ‘enhanced’ component of the harmonic complex. Viemeister and Ba-

con [26] explain their results in terms of adaptation and suppression. The precursor adapts neurones (probably in the cochlear nerve) sensitive to the frequencies in the signal. A test signal with energy at all frequencies will excite the unadapted neurones more strongly than those that have been adapted. At a higher level (cochlear nucleus), neurones corresponding to frequencies in the precursor signal are excited less vigorously than without adaptation and these, in turn, suppress neurones sensitive to neighbouring frequencies less. The result is that the frequency components present in the test signal but not in the precursor are effectively enhanced.

In the present study we examine the effects of prior exposure to noise on the intelligibility of voiced plosives, compared to coincident noise. Low-pass, high-pass, band-pass and band-stop filtered noise has been employed with preceding noise durations of 0 (coincident noise) to 800 ms.

In our previous experiments, it was found that speech could be recognized better in continuous noise than in noise which is switched on at the beginning of the speech and off at the end, and it was suggested that this was due to some adaptation process. In the present study we try to test this hypothesis in a new series of experiments in where the spectrum of the noise is varied in order to investigate adaptation in different frequency regions.

We also attempt to estimate the time scale of the adaptation process. It is expected, from an adaptation process, that it will be faster at the beginning and then, stabilize. Intelligibility might be a linear function of the logarithm of preceding noise duration.

2. Experiments

The objective was to measure the intelligibility of voiced plosives in consonant-vowel syllables in noise as a function of the duration of the noise preceding the syllable.

2.1. Stimuli

A set of 18 syllables were spoken by a male British English speaker and recorded digitally at a sampling rate of 11,025 Hz. The syllables consisted of the voiced plosives /b/, /d/, /g/ combined with six British English vowels /i/, /e/ (as in bean, sea), /e/ (as in bed, dress), /e/ (as in nurse, stir), /a/ (as in barn, start), /a/ (as in born, law), and /a/ (as in boon, goose) [27]. All stimuli were made equal in duration by extracting 100 ms of speech beginning at either the murmur preceding the plosive burst or at the burst itself. The amplitudes were adjusted to make them of equal power across the entire syllable. The amplitude corresponding to the short vowel portion did not differ so much across vowels. The first ten speech samples were multiplied by the first half of a 20 point Hanning window and the last ten points by the second half of the window. Filtered noise was added to make the SNR during the syllable equal to 0 dB rms.

White noise was passed through a digital FIR filter of order 200. Eleven different filters were employed: 1 kHz,
2 kHz and 3 kHz low-pass, 1 kHz, 2 kHz and 3 kHz high-pass, 1–2 kHz and 2–3 kHz band-pass and 1–2 kHz and 2–3 kHz band-stop and all-pass (or 5.5 kHz low-pass as the sampling rate was 11.025 kHz). Figure 1 shows the frequency response of one of the low-pass, band-pass, band-stop, and high-pass filters. The noise began at 0 ms (coincident), 100 ms, 200 ms, 400 ms or 800 ms before the onset of the syllable and continued to the end 100 ms later. The onset and end of the noise were also subjected to the Hanning window as above.

2.2. Listeners

Nine listeners, staff or students of the Universities of Keele or Valencia, took part in the experiments. Six were male and three were female. Their ages ranged from 25 to 59 years. All but two of the listeners were native English
speakers. Of the others, one was Spanish and one Greek but both had a good command of the English language and in the experiments neither obtained scores greater than or less than those of the best or worst English listeners. None of the listeners reported any hearing difficulties.

2.3. Conditions
The experiments took place in a sound-proof room using Sennheiser HD 435 headphones. The stimuli were presented at a level of 70 dB(A). They were controlled by a computer placed external to the sound-proof room. On hearing a stimulus the listener was instructed to press one of three keys labelled 'B', 'D', or 'G'. The next stimulus was presented 2 s after the key had been pressed.

In each session there were 3 (plosives) × 6 (vowels) × 5 (repetitions) = 90 stimuli with the same noise spectrum and duration (0, 100, 200, 400 or 800 ms). The stimuli were presented randomly during a listening session and the sessions with different noise durations were also randomised. Each listener took part in 5 (durations) × 11 (filter conditions) = 55 sessions spread over a few weeks.

3. Results
Each session was scored by counting the number of times each consonant was correctly identified. The scores ranged from near chance (33%) to near perfect (100%) recognition. In practice the highest mean score for any condition was about 90% as no stimuli were presented without the addition of noise and the duration of all syllables was 100 ms, which is rather short.
Table I. F and p values of the ANOVA for each of the plosives separately in low-pass, high-pass, band-pass, and band-stop filtered noise, with noise duration and filter cut-off frequency as factors.

<table>
<thead>
<tr>
<th></th>
<th>Noise duration</th>
<th>Filter cut-off frequency</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-pass</td>
<td>b</td>
<td>F = 4.41 (p &lt; .01)</td>
<td>F = 12.06 (p &lt; .01)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>F = 4.25 (p &lt; .01)</td>
<td>F = 45.77 (p &lt; .01)</td>
</tr>
<tr>
<td></td>
<td>g</td>
<td></td>
<td>F = 13.81 (p &lt; .01)</td>
</tr>
<tr>
<td>High-pass</td>
<td>b</td>
<td>F = 4.54 (p &lt; .01)</td>
<td>F = 28.13 (p &lt; .01)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td></td>
<td>F = 11.1 (p &lt; .01)</td>
</tr>
<tr>
<td></td>
<td>g</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Band-pass</td>
<td>b</td>
<td>F = 34.89 (p &lt; .01)</td>
<td>F = 3.22 (p &lt; .05)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>F = 6.84 (p &lt; .01)</td>
<td>F = 4.41 (p &lt; .01)</td>
</tr>
<tr>
<td></td>
<td>g</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Band-stop</td>
<td>b</td>
<td>F = 28.73 (p &lt; .01)</td>
<td>F = 15.18 (p &lt; .01)</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>F = 49.47 (p &lt; .01)</td>
<td>F = 18.55 (p &lt; .01)</td>
</tr>
<tr>
<td></td>
<td>g</td>
<td></td>
<td>F = 38.31 (p &lt; .01)</td>
</tr>
</tbody>
</table>

3.1. Low-pass

Figure 2 shows the recognition scores and their standard errors as a function of noise duration for the 1, 2, 3 and 5.5 kHz low-pass filters. An ANOVA was carried out with noise duration and filter cut-off frequency as the factors. It was found that the effects on the mean score of both the noise duration ($F_{4,32} = 5.13, p < .01$) and the filter frequency ($F_{3,24} = 72.94, p < .01$) were significant, but not the interaction. The mean score increased with increasing duration but decreased with increasing filter frequency.

Figure 2 suggests that for 1 kHz, 2 kHz and 3 kHz noise there is an increase in recognition score with noise duration, but for the 5.5 kHz noise (effectively white noise at 11,025 Hz sampling rate) there is an initial drop in score at 100 ms, which is similar to that found previously [18]. However, the simplest explanation of the 5.5 kHz results is that there is no effect of duration with wide-band noise. Separate ANOVAs were also carried out for the scores with each of the plosives (Table I). The factors were noise duration and filter cut-off frequency.

3.2. High-pass

The recognition scores and their standard errors as a function of noise duration for the 1, 2 and 3 kHz high-pass filters are shown in Figure 3. An ANOVA was carried out with noise duration and filter cut-off frequency as the factors. It was found that the effects on the mean score of both the noise duration ($F_{4,32} = 3.59, p < .05$) and the filter frequency ($F_{2,16} = 21.83, p < .01$) were significant, but not the interaction. The mean score increased with both increasing duration and filter frequency.

As with the low-pass filters, ANOVAs were also carried out for the scores with each of the plosives separately. The factors were noise duration and filter cut-off frequency. The results are also shown in Table I.

3.3. Band-pass

The recognition scores and their standard errors as a function of noise duration for the 1–2 and 2–3 band-pass filters are shown in the top two panels of Figure 4. An ANOVA, which included the 1 kHz low-pass (0–1 kHz band-pass) and 3 kHz high-pass (3–5.5 kHz band-pass) scores was carried out with noise duration and filter pass-band frequency as the factors. It was found that the effects on the mean score of both the noise duration ($F_{4,32} = 21.39, p < .01$) and the filter frequency ($F_{2,16} = 12.09, p < .01$) were significant, and so was the interaction ($F_{12,996} = 2.96, p < .01$). The mean score increased with increasing duration but fell with increasing filter frequency to a minimum with the 1–2 kHz filter, then rose again as the pass-band frequency further increased.

Figure 4 shows that the increase in score between the 0 ms (co-gated) condition and the 800 ms condition for the 1–2 kHz band-pass noise is about 15%. This is similar to the result found in the original experiments with speech-shaped noise [7].

ANOVA was also carried out for the band-pass scores with each of the plosives separately (Table I). The factors were noise duration and filter cut-off frequency.

3.4. Band-stop

The recognition scores and their standard errors as a function of noise duration for the 1–2 and 2–3 band-stop filters are shown in the bottom two panels of Figure 4. An ANOVA, which included the 1 kHz high-pass (0–1 kHz band-stop) and 3 kHz low-pass (3–5.5 kHz band stop) scores were carried out with noise duration and filter pass-band frequency as the factors. It was found that the effects on the mean score of both the noise duration ($F_{4,32} = 22.97, p < .01$) and the filter frequency ($F_{2,16} = 46.86, p < .01$) were significant, and so was the interaction ($F_{12,996} = 2.04, p < .05$). The mean score increased with increasing duration but also rose with increasing filter frequency to a maximum with the 1–2 kHz band-stop filter, then fell as the stop-band frequency further increased.

ANOVA was also carried out for the band-stop scores with each of the plosives separately (Table I).
4. Acoustic analysis

As the duration and spectral shape of the preceding noise had a significant effect on the score for some plosives but not others, a spectral analysis of the plosives was carried out so that the effects of the noise on the features of the plosives could be examined. It is known [28] that the spectrum of the plosive bursts and the direction and extent of the formant transitions are important perceptual features for distinguishing between the plosives. Blumstein et al. [29] suggested that place of articulation is derived from a set of integrated properties characterised by the short-term spectrum at the point of release and a set of simple properties such as burst frequency. Furthermore Blumstein and Stevens [30] suggested that properties derived from this onset spectrum are context invariant. However, the exact nature of the integration is not yet known. Dorman and Loizou [31] found that, although relative spectral change could be successfully employed for separating labial from alveolar voiced plosives, it was the formant transitions which were used as perceptual cues.

In order to examine these features, linear prediction spectra were computed at the burst of each plosive and at approximately half way through each stimulus. Each plo-
sive burst was analysed by computing a 10th order LPC spectrum with a 128-point (11.6 ms) window. Ten coefficients were chosen, as the bandwidth of the signals was limited to 5.5 kHz (twice the number of formants (four) plus two). These spectra are shown as the solid lines in Figures 5, 6 and 7. LPC spectra were also computed 512 points (45.4 ms) from the start of each stimulus. These are shown as the dashed lines in Figures 5, 6 and 7. For clarity, these are plotted displaced by 20 dB on the ordinate from the burst spectra.

It can be seen from Figure 5 that the burst of /be/ has a main peak at about 300 Hz, that of /de/ at about 3600 Hz, and that of /ge/ at about 2200 Hz. For /ba/ the main peak is again at about 300 Hz but those of /du/ and /gu/ are lower, with that of /du/ at about 2800 Hz and that of /gu/ at 1400 Hz. Figure 6 shows that the /o/ syllables have a similar pattern, with the burst of /bo/ at about 300 Hz, that of /do/ at about 2700 Hz and that of /go/ at about 1200 Hz. Figure 6 also shows that the structure of some of the bursts of the /u/ syllables are more complex; /ba/ has
a peak at about 300 Hz but also another at about 1300 Hz. The largest peak of /da/ is at 1800 Hz but there are others at higher frequencies, particularly at about 3400 Hz. For /ga/ there is a prominent single peak at about 1600 Hz. Figure 7 shows that some of the /i/ and /e/ syllables also have complex bursts. The main peak of /be/ is at about 400 Hz and /ge/ shows a broad peak around 300 Hz, but /de/ has three peaks distributed throughout the spectrum. With the /i/ syllables, /bi/ also has three peaks of similar heights, /di/ has a fairly flat spectrum up to about 3600 Hz, and /gi/ has a peak at about 250 Hz and a broad peak around 3000 Hz.

As the glottis begins vibrating at or before the release in voiced plosives in English, the spectra shown by solid lines in Figures 5, 6 and 7 represent both the energy of the burst and that of the first one or two glottal pulses. Some of the peaks in the spectra thus represent formant frequencies and others the release burst. The peaks in the dashed curves show the formant frequencies some 45 ms later. From the two spectra the time course of the formant transitions can be estimated.

5. Perceptual analysis of individual plosives

Acoustic analysis shows which features are available for discriminating speech sounds but only perceptual analysis can reveal which features are actually used. The perceptual analysis consisted of the calculation of correct mean scores, and construction of confusion matrices. As there were 330 confusion matrices (6 vowels × 11 filtered noise conditions × 5 noise duration conditions) they are not presented and only those data that are relevant to explain some of the results are discussed.

A combination of the results of the perceptual experiments with different noise conditions can show which regions of the spectrum are used for distinguishing between different plosives. Furthermore, the results can also show which features are unmasked by preceding noise.

The results in Figures 2, 3 and 4 show that perception of the syllables depends upon both the duration and the spectrum of the preceding noise. As a way of smoothing the data, linear regression lines were fitted to the scores but using a logarithmic time axis (assuming that 0 represents coved noise cognition). It might be expected, for an adaptation process, that intelligibility would be a linear function of the logarithm of noise duration, so straight lines should fit the data approximately. The intercept of these lines is a measure of the intelligibility when the noise is coincident with the speech, and the slope is a measure of the amount of adaptation. The correct scores for the four band-stop conditions (1 kHz high-pass, 1–2 and 2–3 band-stop and 3 kHz low-pass) show the approximate frequency regions where the perceptual features reside, and the scores for the four band-pass conditions (1 kHz low-pass, 1–2 and 2–3 band-pass and 3 kHz high-pass) show complementary information. However, it must be remembered that, although the noise may mask a feature, for recognition purposes the presence or absence of a feature in the region of the noise is rendered unknown ('missing' data, [32]).

Figures 8, 9 and 10 show the scores for /b/, /d/ and /g/ for each of the vowels. The solid lines show the scores for the 0–1, 1–2, 2–3 and 3–5.5 kHz band-pass filters and the dashed lines for the 0–1, 1–2, 2–3 and 3–5.5 kHz band-stop filters for the intercepts of the regression lines (noise coincident with the speech). The one-sided 'error' bars show the change in scores for 400 ms of preceding noise (derived from the slope of the regression lines). The duration of 400 ms was chosen arbitrarily to show changes large enough to be seen but not so large that a score of 100% was exceeded.
For the /l/ syllables (Figure 8) the scores with all of the band-stop filters show a peak in the 1–2 kHz region, suggesting that sufficient information for discriminating the plosives lies in this region. With the band-pass filters, noise above 3 kHz has little effect on the /l/ and /d/ scores, but the /g/ score drops, presumably because the falling spectrum of the /g/ burst in this region is changed with the presence of the coincident noise so that it sounds like the rising spectrum of the /d/. For a total of around 30% of errors in the recognition of /g/ /28% corresponded to confusions with /d/.

The scores for the /u/ syllables are also shown in Figure 8. For the band-stop filters there is a peak in the low frequency region for /u/ where the burst energy lies. For /u/ there is a peak in the 1–2 kHz region. Figure 5 shows that the burst spectrum of /u/ has a dip this region. But it can be seen that there is a peak of energy in the spectrum obtained at 45 ms from the onset of the stimulus (dashed line), suggesting that the recognition scores of this syllable may be based on the second formant transition. Figure 5 shows that the burst spectrum of /u/ has a dip in this region where /u/ and /g/ have peaks. For the band-pass filters the score for /u/ is low when the noise is below 2 kHz, masking the peaks, but higher above this where there is no significant energy. The scores for /u/ are low with noise in the 1–2 kHz region. This suggests that the second formant transition is masked, or that the presence of band-pass noise in this region allows confusion with /g/, which also has a peak of energy in the 1–2 kHz frequency region (for a total of around 48% of errors in /u/ perception, 39% corresponded to confusions of /u/ with /g/); also for /g/, noise in the 2–3 kHz regions allows confusion with /u/ (for a total of around 28% of errors in /u/ perception, 21% corresponded to confusions of /u/ with /g/).

Figure 9 shows the scores for the /l/ syllables. For /b/ the band-stop noise gave the highest score for the 0–1 kHz region corresponding to the peak in the burst spectrum. For /d/ and /g/ the best recognition was in the 1–2 kHz region. With no noise in this region (band- stop 1–2 kHz) the peak in the /g/ burst can be heard, enabling it to be distinguished from /d/. For the band-pass filters, noise below 2 kHz depressed the /b/ scores, in the 1–2 kHz region lowered the /d/ scores and in the 2–3 region reduced the /g/ scores.

The results for /u/ shown in Figure 9, are difficult to explain as band-pass and band-stop noise show similar effects except in the region above 3 kHz. Adding noise in this region increases the score but adding noise below this region does not. For /u/ the effects of band-pass and band-stop noise are complementary, with a peak in the scores with no noise between 1 and 2 kHz and a dip with noise in that region. This noise produces more confusions with /g/ which also shows a peak in the 1–2 kHz region. No noise in the 2–3 kHz region, corresponding to the dip in the burst spectrum (Figure 6), increases the /g/ score whereas adding noise in this region slightly lowers it.

Figure 10 shows that the /l/ score is best with noise only above 3 kHz (band-pass 3–5.5 kHz) or by having no noise below 1 kHz. The /l/ score is reduced by adding noise in the 1–2 kHz region, in which its spectrum presents a peak, and increases by adding noise everywhere except in this region. The spectra (Figure 7) show that the information about the identity of the plosive is probably distributed throughout the spectrum. Adding noise only above 3 kHz or below 1 kHz increases the scores of all the
/ge/ syllables, suggesting that formant transitions rather than the burst spectra are important.

The scores for the /bi/ syllables are also shown in Figure 10. Low frequency band-stop noise produces the highest /bi/ scores, medium frequency band-stop noise the highest /gi/ scores and higher frequency band-stop noise the highest /di/ scores, corresponding to the positions of the peaks in the burst spectra (Figure 7). Adding noise below 2 kHz lowered the /bi/ scores, in the 2–3 kHz region lowered the /di/ scores, but above 3 kHz raised the /gi/ scores.

6. Noise adaptation

It was seen in Figures 2, 3 and 4 that when the noise preceded the syllable the score generally increased. The extent of the improvement in recognition due to preceding noise, compared to coincident noise, for each syllable, is shown by the bars added to the scores in Figures 8, 9 and 10. These were computed by using the regression lines to the data shown in Figures 2, 3 and 4. The slope of these lines gives a measure of the adaptation caused by the preceding noise. In Figures 8, 9, and 10 the slope is converted into the estimated change in score, which would be produced by the noise leading the speech by 400 ms. The end of each bar is thus the score which would be expected in this condition.

For each condition, each syllable was heard 5 times with 5 preceding noise durations by each of the 9 listeners. An ANOVA of the scores was computed for each condition with noise duration and listener as the factors. The significance of duration is shown in Figures 8, 9 and 10 by * (p < .05) and ** (p < .01) attached to the bars.

For /be/ only the 2–3 kHz band-stop (F_{4,32} = 4.91, p < .01) was significant but for /ge/, the band-pass 1–2 kHz (F_{4,32} = 4.43, p < .01) and >3 kHz (F_{4,32} = 21.16, p < .01) filters and the band-stop 2–3 kHz (F_{4,32} = 5.85, p < .01) filter were significant (Figure 8). For /de/, no filters were significant. For /bu/, the band-pass 1–2 kHz (F_{4,32} = 11.34, p < .01) and the band-stop 2–3 kHz (F_{4,32} = 3.70, p < .05) filters were significant and for /gu/, the band-stop 2–3 (F_{4,32} = 4.55, p < .01) and >3 kHz (F_{4,32} = 4.84, p < .01) filters were significant (Figure 8). For /di/, no filters were significant.

For /bo/, the band-stop 1–2 kHz (F_{4,32} = 81.0, p < .01), 2–3 kHz (F_{4,32} = 12.63, p < .01), >3 kHz (F_{4,32} = 2.87, p < .05) and the band-pass 1–2 kHz (F_{4,32} = 8.12, p < .01) filters were significant (Figure 9). For /go/, only the band-stop 2–3 kHz filter was significant (F_{4,32} = 25.69, p < .01). For /gu/, the band-pass 0–1 kHz (F_{4,32} = 4.36, p < .01) and 2–3 kHz (F_{4,32} = 6.64, p < .05) filters were significant. For /bu/, only the band-stop >3 kHz (F_{4,32} = 5.90, p < .01) was significant. For /di/, the band-pass 0–1 kHz (F_{4,32} = 2.93, p < .01) and the band-stop 2–3 kHz (F_{4,32} = 3.38, p < .05) and >3 kHz (F_{4,32} = 4.70, p < .01) filters were significant. For /ge/, the band-stop 1–2 kHz (F_{4,32} = 9.92, p < .05), 2–3 kHz (F_{4,32} = 2.80, p < .05) and >3 kHz (F_{4,32} = 5.07, p < .01) were significant (Figure 9).

For /be/, the band-pass 0–1 kHz (F_{4,32} = 3.59, p < .05) filter and the band-stop 1–2 kHz (F_{4,32} = 4.59, p < .01) and 2–3 kHz (F_{4,32} = 5.44, p < .01) filters were significant, and so was the band-pass 1–2 kHz (F_{4,32} = 8.70, p < .01) for /de/. None were significant for /ge/. For /bi/, the band-pass 0–1 kHz (F_{4,32} = 5.46, p < .01) filter and the band-stop 2–3 kHz (F_{4,32} = 11.34, p < .01) filter were significant. For /di/, the band-stop 2–3 kHz (F_{4,32} =
3.28, $p < .05$) and $>3$ kHz ($F_{3,32} = 3.80, p < .05$) filter were significant. For /g/ı, the band-stop 1–2 kHz ($F_{3,32} = 3.14, p < .05$) and 2–3 kHz ($F_{3,32} = 3.69, p < .05$) filters were significant (Figure 10).

7. Discussion

Experiments have been carried out to determine the effects of preceding filtered noise on the perception of voiced plosives. It has been confirmed that their intelligibility in CV syllables increases as the duration of the noise is lengthened but the amount of increase depends upon the spectral shape of the noise employed. It appears that band-stop noise is most effective, band-pass next, then low-pass, with high-pass being least effective. It seems that the presentation of a preceding narrow band of noise (two better than one) produces more beneficial effects on the recognition of the plosives than low-pass or high-pass noise.

There are a number of mechanisms which could be responsible for the improvement in the recognition of plosive-vowel syllables in continuous preceding noise, compared with coincident noise, found in the present experiment. One possible mechanism is the raising of firing thresholds of neurons in the cochlear nerve [8, 9, 10]. It would be expected that this would only take place in neurons whose receptive fields overlap the frequency of the noise. Raising the threshold and lowering the spike discharge rate effectively subtracts energy from the signal in this frequency region. A model based on this mechanism was developed by Ainsworth and Meyer [7].

Another possibility is a more central mechanism such as auditory streaming [33]. Components of the auditory ‘scene’, which share common attributes such as fundamental frequency or onset time, are assumed to emanate from the same source and are grouped together. In the present experiments, with noise preceding the speech, syllables might be assumed to originate from a separate source from the speech and so will be assigned to a different auditory stream; whereas noise which is coincident, will remain in the same stream as the speech and will be included in any pattern matching process. This will lead to a higher recognition score for the stimuli with preceding noise.

Both peripheral adaptation and central streaming lead to a mechanism functionally equivalent to spectral subtraction [34]. However, it might be expected that peripheral adaptation would take time to build up, leading to a gradual increase in recognition score with duration of preceding noise; whereas as soon as the preceding noise is established as a separate auditory source, central streaming predicts that the score will increase sharply. The present experimental results support both hypotheses equally. From Figures 2, 3, and 4, five filters (1 and 3 kHz low-pass, 1 kHz high-pass, 1–2 kHz band-pass and 1–2 kHz band-stop) show a gradual build up and five (2 kHz low-pass, 2 and 3 kHz high-pass, 2–3 kHz band-pass and 2–3 kHz band-stop) show a more abrupt change.

Both mechanisms lead to the hypothesis that preceding noise will reduce the masking effect of that noise, compared with coincident noise, in that part of the spectrum where the noise occurs. From Figure 5 it can be seen that the burst spectrum of /d/ı has a peak at about 3500 Hz and that of /g/ı has one at about 2200 Hz. Adding 3–5.5 kHz band-pass coincident noise causes the /g/ı spectrum to look more like the /d/ı spectrum, suppressing the /g/ı score (Figure 8). Adding preceding noise instead, effectively reduces the effect of the noise, leading to a large increase in the /g/ı score. A similar effect is produced by the preceding 2–3 kHz band-stop noise (Figure 8).

Coincident noise in the 1–2 kHz region causes /b/ı to be confused with /g/ı, as it makes its burst spectrum appear more like that of /g/ı, which has a peak in this region (Figure 5). Preceding noise in this region reduces the effect, producing an increase in the /b/ı score (Figure 8).

The effects of adding preceding noise are not universally beneficial and it is greater with some syllables than with others. Figure 9 shows that preceding band-stop 2–3 kHz noise increases the /b/ı score but reduces the /d/ı score compared to that with coincident noise. Spectral subtraction of the noise increases the discrimination of /b/ı and /g/ı, which have peaks at about 500 and 1200 Hz respectively (Figure 6), but produces a rather flat /d/ı spectrum. Consequently the /b/ı score increases but /d/ı becomes ambiguous and its score declines. A similar situation occurs with 3 kHz low pass noise and the /a/ syllables (Figure 9). Preceding noise enables the /b/ı and /g/ı/ı peaks to become more prominent (Figure 6), leading to increased scores for those syllables, whilst the /d/ı/ı spectrum is flattened and its score declines.

Little can be deduced from the /le/ı scores (Figure 10) as all the significant increases due to preceding noise occur with /b/ı/ı. Preceding 2–3 kHz band-stop noise causes increases in the /b/ı/ı and /d/ı/ı scores, and a (non significant) fall in the /d/ı/ı score (Figure 10). It can be seen from Figure 7 that the second formant (F2) transitions (see the burst spectrum (solid curve) and the vowel spectrum (dashed curve)) for /b/ı/ı (1500 Hz to 2600 Hz) and /d/ı/ı (1700 Hz to 2600 Hz) rise and cross the boundary of this region, whilst that of /g/ı/ı (2800 Hz to 2600 Hz) falls but remains within the 2–3 kHz region. Hence reducing the effect of the noise reduces the disruption to the F2 transitions of /b/ı/ı and /d/ı/ı, leading to increased scores (Figure 10).

It is possible that the increased recognition scores with longer durations of preceding noise is related to the psychoacoustical ‘enhancement’ effect [19, 20, 21, 22, 25, 23]. Several explanations of the enhancement effect involving attention and adaptation have been hypothesised. The results of experiments by Viemeister and Bacon [26] favour an ‘adaptation’ explanation. They explain their results in terms of adaptation and suppression. Such a mechanism could be responsible for the results of the present experiments. With preceding ‘white’ noise all frequencies will be adapted equally so there will be less effect. With band-stop noise the frequencies in the syllables within the stop-band will be enhanced, leading to higher recognition.
scores when the acoustic cues for recognition lie within this frequency band. With low-pass and high-pass noise, a region above or below, respectively, the cut-off frequency of the filter will be enhanced, and with band-pass noise two regions below and above the pass-band will be enhanced. Hence the effect of preceding noise will be greatest for band-stop noise, less for band-pass noise, least for low-and high-pass noise and non-existent for white noise. This is approximately what was found.

8. Conclusions

It has been shown that the recognition scores of voiced plosives in consonant-vowel syllables with added noise increase with the duration of the noise preceding the syllable. For a duration of 800 ms this increase was about 14% for band-stop noise, 12% for band-pass noise, 7% for low-pass noise and only 3% for high-pass noise.

The phenomenon could be due to a peripheral adaptation mechanism or to a central mechanism, such as auditory scene analysis (or a combination of the two). A mechanism based on adaptation in the cochlear nerve affecting the suppression of inhibition in the cochlear nucleus is consistent with the results obtained. The result of this mechanism is that preceding noise leads to an effective reduction in noise level in the region of the spectrum where the noise occurs and enhancement of the signal in the other parts of the spectrum. This phenomenon probably makes human speech recognition more robust in noisy environments.

Acknowledgement

This work was supported by an EU TMR contract No. FMRX-CT97-0150.

References


