© S. Hirzel Verlag
Stuttgart
ACUSTICA
Vol. 78 (1993)

The Perception of Speech in Fluctuating Noise

P. A. Howard-Jones, S. Rosen
Department of Phonetics & Linguistics, University College London

The Perception of Speech in Fluctuating Noise

Summary

Various phenomena associated with the perception of speech in fluctuating noise were investigated using a small set of VCV (Vowel-Consonant-Vowel) stimuli. In experiments 1 and 2, the performance intensity function and the speech reception threshold were measured for the speech in noise backgrounds which fluctuated on various time-scales. It was found that very rapid fluctuations in the masker had no effect on speech perception, whilst a masker containing fluctuations of medium duration (80 to 100's of milliseconds) provided a release from masking but did not affect the relative intelligibility of each consonant, nor the pattern of confusions. A noise environment containing longer interruptions (of duration 1 second) produced a flattening of the performance intensity function and a randomisation of the types of phonetic confusion occurring. In experiments 3 and 4, the effect on the speech reception threshold of varying the duration and magnitude of fluctuations was studied. It was found that the ability of subjects to "glimpse" speech information, which may be used to explain the release of masking, depends strongly upon both the duration of the quiet sections of the noise and also upon their magnitude. A method for quantifying the fluctuation of a speech masker is proposed and examples of the analysis are given for cafeteria noise, and white noise shaped to possess similar frequency content. The analysis shows no differences in the fluctuation of the two noises likely to influence speech intelligibility, a result confirmed by experiment.

Die Wahrnehmung von gesprochener Sprache bei fluktuierendem Störgeräusch

Zusammenfassung

Eine Reihe von Beobachtungen, die mit der Wahrnehmung von gesprochener Sprache bei variierendem Lärm zusammenhängen, werden unter Zuhilfenahme eines Satzes von VDV-Stimuli(Vokal-Konsonant-Vokal-Stimuli) untersucht. Im ersten und zweiten Experiment werden die PI-Funktionen (Performanz-Intensitäts-Funktionen) und die SRT (*s*peech *r*eception *t*hreshold: Sprachverständlichkeitsgrenze) für gesprochene Sprache auf einem Lärmhintergrund gemessen, der sich mit der Zeit ändert. Es stellt sich heraus, daß sehr rasch aufeinanderfolgende Variationen des Maskierers keinen Einfluß auf die Sprachverständlichkeit haben, ein Maskierer mit einer Periodendauer von mittlerer Länge (80–100 Millisekunden) aber eine Verdeckungsminderung ergibt, die jedoch die relative Verständlichkeit der verschiedenen Konsonanten und das Muster der vorkommenden Verwechslungen nicht beeinflußt. Eine Lärmumgebung mit längeren Unterbrechungen (von einer Sekunde Dauer) bewirkt eine Abflachung der PI-Funktion und die Verwechslungen werden unvorhersagbar. In den Experimenten 3 und 4 wurde untersucht, welchen Einfluß unterschiedliche Dauer und Größe der Fluktuationen auf die SRT haben. Es stellt sich heraus, daß die Fähigkeit von Versuchspersonen, Redeinformation zu „erhaschen" (die die Verdeckungsminderung erklären könnte) stark von der Dauer und der Größe der stillen Perioden des Lärms abhängt. Es wird eine Methode vorgestellt, anhand derer die Fluktuationen des Sprachmaskierers gemessen werden können. Die Analyse erfolgt anhand von zwei Beispielen: zum einen mit Umgebungsgeräuschen aus einer Cafeteria, zum anderen mit weißem Rauschen, dessen Langzeitspektrum dem Langzeitspektrum des Cafeteria-Lärms angepaßt ist. Die Analyse zeigt, daß die unterschiedlichen Fluktuationen der beiden Lärmarten, welche die Sprachverständlichkeit beeinflussen könnten, keine Änderung bewirken; ein Ergebnis, das durch Versuche bestätigt wird.

La perception de la parole dans un bruit variable

Sommaire

Nous avons étudié divers phénomènes associés à la perception de la parole en présence de bruit variable, sur un jeu limité de stimuli VCV. Dans les deux premières expériences, nous avons mesuré la fonction PI (Performance Intensity) et le SRT (Speech Reception Threshold) pour évaluer la parole dégradée par un bruit masquant d'intensité variable au cours du temps, à différents pas. Nous avons observé que de très rapides fluctuations du masque n'ont eu aucun effet sur la perception de la parole. En revanche, un masque fluctuant sur des durées moyennes (de 80 à quelques centaines de millisecondes) provoque une atténuation du masquage, san affecter pour autant l'intelligibilité relative de chaque consonne, ni la structure des confusions observées. Un environnement bruité contenant des interruptions plus longues (de l'ordre de la seconde) produit un aplatissement de la fonction PI et rend aléatoires les confusions phonétiques résultantes.

Dans les deux dernières expériences, nous avons étudié l'effet de la modulation du bruit en durée et en amplitude sur le SRT. Nous avons constaté que la capacité des sujets à «entr'apercevoir» l'information de parole – qui peut justifier la réduction du masquage – est étroitement liée à la durée des portions où le bruit est atténué qui'à leur amplitude.

Nous proposons une méthode pour quantifier la fluctuation d'un bruit de masquage de la parole et nous donnons des exemples basés sur des bruits de cafétéria et sur du bruit blanc déformé de façon à présenter un spectre similaire. L'analyse ne révèle pas de différences susceptibles d'influencer l'intelligibilité de la parole dans la fluctuation des deux bruits; ce résultat est confirmé expérimentalement.

Received 26 March 1992,
accepted 15 December 1992.

P. A. Howard-Jones, S. Rosen, Dept. of Phonetics & Linguistics, University College London, 4 Stephenson Way, London NW1 2HE, England.

1. Introduction

Amongst the extensive literature concerned with speech perception in noise, an early study by French and Steinberg [9], who used filtered white noise, clearly demonstrated that the level and frequency content of a background noise greatly influence its masking effectiveness. Many studies of speech intelligibility, such as those concerned with the effects of hearing impairment, have since restricted themselves to using white noise as a conveniently replicated laboratory standard (e.g. Hawkins and Stevens [12]; Keith and Talis [16]), or a masker created by filtering white noise until it displays spectral characteristics similar to the long-term spectrum of a speech corpus (e.g. Foster and Haggard [8]; Bosman [1]). Other workers, however, have preferred to use speech babble (e.g. Gordon-Salant [10]; Kalikow and Stevens [15]) or cafeteria noise (e.g. Cooper and Cutts [2]; Dirks et al. [4]) on the basis that such noises more closely represent the types of noise environment encountered in everyday situations. An important characteristic of such "real" noises is their fluctuation in overall power.

Despite the general realization that everyday noise environments fluctuate, the effects of fluctuation on speech perception have received far less attention than the effects of the frequency content of the masker and the signal-to-noise ratio. Two early studies (Miller [18]; Miller and Licklider [19]) investigated the effects on speech perception of an interrupted broadband masker. Speech scores were measured for monosyllables in noises that were interrupted at various rates with various noise-time fractions. They observed that, when compared at similar rms values, interrupted noise was often unable to mask speech as well as continuous noise, and this release of masking was greatest when the rate of interruption was around 10 Hz. They explained: "When there were 10 bursts of noise per second, the listeners were able to get several glimpses of every word and to patch these glimpses together...". The magnitude of this effect was not greatly affected when the regularly interrupted noise was replaced by one containing randomly occurring interruptions. In short, for the on/off type of interruptions used in the study, it was concluded that the degree of masking release was highly dependent on the duration of the quiet sections.

More recently, the intelligibility of sentences masked by fluctuating noise has been studied for hearing impaired and normally hearing subjects by Festen and Plomp [6]. Their study showed how the release of masking observed for fluctuating noise is significantly reduced amongst the hearing impaired, and related the ability to "glimpse" speech information with the temporal acuity of the listener. This investigation also reported that a significant release of masking may be observed when the masker is a competing speech signal, thus demonstrating that such effects may commonly occur in everyday environments and are not restricted to laboratory conditions. In an earlier study that also included competing speech signals (Speaks et al. [25]), it was reported that the performance intensity function was flatter than that observed for continuous noise, and this was believed to be due to fluctuations in the power of the masking speech. It is thought that such fluctuations may introduce an extra element of randomness into the perceptual task, making successful identification less dependent on level and more upon the chance occurrence of the utterance during a quiet section of the noise. This result has an important implication for anyone trying to improve the clarity of a speech signal by limiting a background interference which is fluctuating, since increasing the signal-to-noise ratio may give only a disappointing increase in intelligibility.

There is already enough evidence then, to demonstrate the inadequacy of characterising a noise background simply in terms of its average level and frequency content. It is not clear, however, how masker fluctuation should be included in such analyses. To characterise completely the fluctuation of a signal, it is necessary to describe exhaustively how the energy of that signal varies with time. Various methods are already available in other scientific fields for partially describing the time-variation of a signal, but these simple measures were not formulated to include those dimensions of fluctuation that are relevant to the masking of speech. For example, waveform moments have been used in psychoacoustical studies involving the perception of sinusoids (Hartmann and Pumplin [11]), and there are many simpler measures such as variance and crest factor often used in other types of study. Yet, these methods do not take into account the duration of the quieter sections, which we already know influences the speech-masking properties of the noise.

Sotscheck [23, 24] used a more complex method in which cumulative distribution functions were constructed from sampling the rectified and smoothed outputs of 1/3-octave filters. Although such an analysis was useful in showing that some maskers (e.g., speech) had much greater fluctuations (and lower masking effectiveness) than others (e.g., pink noise), the parameters controlling the analysis (in particular the time constant of smoothing) do not appear to be chosen on the basis either of empirical effects of masker fluctuation, nor on the properties of the auditory system. It would therefore be difficult to imagine such an analysis being useful in predicting the effects of a masker (nor did Sotscheck attempt this).

Before a more suitable measure of masker fluctuation can be provided, more needs to be known about those issues specific to this type of perceptual task. The present study is concerned with a further investigation of the effects of masker fluctuation, and the first two experiments include an exploration of the associated phenomena at a segmental level. Experiments 3 and 4 were designed to investigate those dimensions of fluctuation assumed pertinent to speech perception, and thus to provide a basis for quantifying the fluctuation of everyday noise environments in terms related to speech intelligibility.

2. Experiment 1

The first experiment attempted to reproduce some of the effects of masker fluctuation observed in previous studies. In this investigation, however, it was also possible to compare the intelligibility scores of individual sounds in different noises of similar frequency content, at signal to noise (S/N) levels resulting in similar overall performance.

2.1. Method

2.1.1. Subjects

In all the experiments to be reported in this paper, subjects were native speakers of English with no known history of hearing disorder. Both males and females participated. The subjects had an age range of 18–30 years, and most were undergraduate students. Here, the subjects were in two groups, one for each part of the experiment, each consisting of 6 volunteers.

2.1.2. Stimuli

The speech material used in these tests was restricted to one token each of twelve intervocalic consonants [m b p v f n d t z s g k] as in [*à* m*a*], spoken by a female with a standard Southern British English accent. Data was sampled at 12.8 kHz.

Three samples of masking noise were used. The first was a 4 second sample of white noise (Fig. 1 a) sampled at 12.8 kHz. The second was 4 seconds of "cubed" white noise (Fig. 1 b). This noise was created by raising the instantaneous amplitude of the white noise sample to the power three, and then scaling the signal down. Such a noise signal retains its flat long-term power spectrum, but is more "peaky" (on an instantaneous time-scale) than the white noise from which it is derived, and sounds like a "crackling" noise background. Fig. 1 d shows a detail of the cubed noise waveform compared with the white noise waveform that was used to produce it. The third noise was also

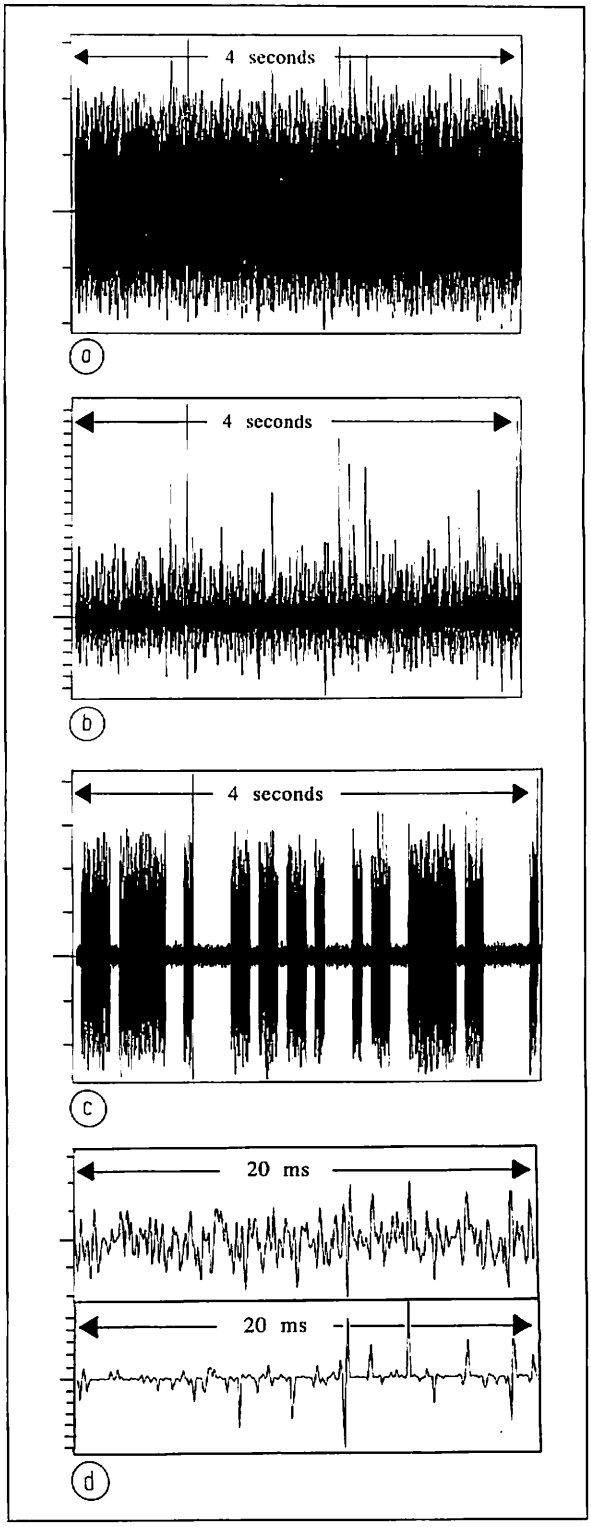

Fig. 1. Noise samples used in experiment 1: a) the white noise sample, b) the "cubed" noise sample (see text), c) the noise sample with interruptions (of durations 80–400 ms), d) Detail of the white noise sample (above) and the cubed noise sample (below) which was derived from it. (All the marks on the abscissae in Figs. 1 and 5 refer to the same arbitrary amplitude scale.)

created digitally, by treating a copy of the original white noise sample in 80 ms time slices, a random selection of which (probability = 0.5) was attenuated by 20 dB (see Fig. 1 c). The signal was scaled in such a way as to possess the same rms energy as the original signal. Again, the processed signal (of 4 seconds duration) exhibited a flat frequency spectrum but contained interruptions with durations and spacings in the range 80 to 400 milliseconds.

2.2. Procedure

2.2.1. Determination of the speech reception threshold

An adaptive "up-down" procedure (Levitt [17]) was used to determine the speech reception threshold. The procedure was controlled by a MASSCOMP computer which played the speech (with noise) at 12.8 kHz (via a 5.6 kHz lowpass filter) through a pair of Sennheiser HD414SL headphones to a subject sitting in a quiet room. The subject was requested to enter his responses via a keyboard. Before testing began, the level of the speech was set such that all the experiments could be performed at a comfortable overall level. During the test, each speech token, selected according to a randomised sequence, was mixed with the noise. The S/N ratio was controlled by fixing the speech level and, before outputting each VCV, scaling the noise sample prior to mixing it with the speech. Only one sample of each type of noise was used in these experiments, but each time that the speech was mixed with the appropriately attenuated noise, a random start point in the 4 second noise file was chosen for the speech data. The level of the noise was increased by 0.5 dB when the subject selected the correct consonant from the list of twelve, and it was reduced by 0.5 dB when an incorrect response was given. Results from a pilot test were used to derive initial S/N levels appropriate for each type of noise.

Every test used each of the 12 consonants four times, randomly arranged into a sequence of 48 trials. Each of 6 subjects was asked to perform two tests for each of the 3 noises (6 tests for each of the 6 subjects). The tests were arranged such that each subject experienced one of the six possible permutations of presentation order twice. A value for the speech reception threshold was derived from every test by calculating the average of the last 36 presentation levels.

2.2.2. Determination of the performance intensity function

This was carried out using the same arrangement of hardware as outlined above. Here, however, each test was carried out at a constant S/N ratio, and the percentage correct score was recorded. Six new subjects each performed six tests with each of the three noises at +12, +6, +3, 0, −3 and −6 dB relative to a previously estimated speech reception threshold. The three batches of six tests were arranged amongst the subjects such that each subject experienced one of the six possible permutations of presentation order.

2.3. Results and discussion

The mean speech reception thresholds for each noise and experimental procedure are given in Table I, as a function of masker noise. In addition, the performance intensity functions were also averaged across subjects (within noise type) to produce the mean performance intensity functions shown in Fig. 2. (All S/N ratios reported in experiments 1 and 2 are relative to the speech reception threshold adaptively measured for white noise in Experiment 1).

Note that speech reception thresholds estimated adaptively were always slightly higher than those estimated from the performance intensity functions. Such differences may arise, of course, from individual differences between the subjects in the two groups. However, they are more likely the results of practice effects: measurement of the performance intensity function allows the subject to gain more familiarity with the test material, so performance measured by this method generally appears better.

An analysis of variance on the adaptively measured speech reception thresholds showed them to differ significantly across noise maskers ($p<0.0001$). Further analysis (using Tukey's H.S.D.) showed that speech reception thresholds for white and cubed noise were not different from one another, whereas that for interrupted noise was significantly different from both the other noises.

Table I. Speech reception thresholds (SRTs) measured for three different masking noises using two different techniques, expressed relative to the SRT for white noise measured adaptively. Adaptive SRTs are the mean of two tests from each of 6 subjects, while performance intensity (PI) estimates were obtained by taking the mean SRT for each subject determined from a logistic fit to the 6 individual PI functions.

Noise type	Adaptive		PI	
	SRT dB	Standard error dB	SRT dB	Standard error dB
Continuous white	0.00	±0.76	−0.49	±0.94
"Cubed"	+0.45	±0.47	−1.65	±0.78
Interrupted (80–400 ms durations)	−4.21	±0.92	−6.45	±1.02

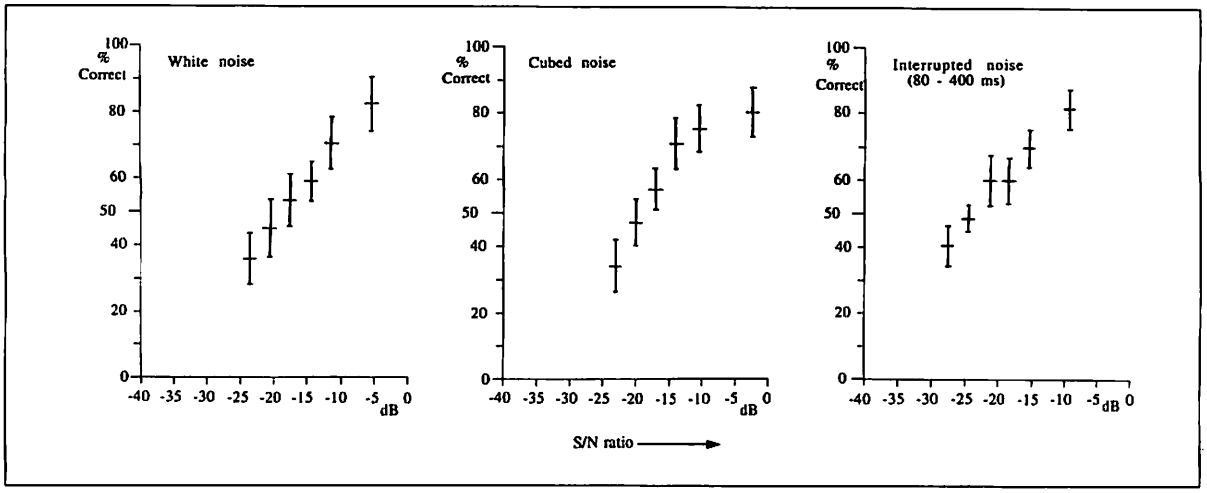

Fig. 2. The performance intensity functions derived from combining results for the six subjects listening to the test material masked by the three noise samples used in experiment 1, presented over a range of S/N levels. The abscissae refer to a common but arbitrary scale of S/N ratio, also used in Fig. 6. (All error bars here and in other figures refer to estimated standard errors in the mean.)

In order to quantify differences among the three performance intensity functions, a statistical approach based on generalized linear models (GLMs) was used. This technique is analogous to the analysis of (co)variance and multiple linear regression for Gaussian distributed data, and can be considered a variant of logistic regression. The predictor variables used in the analysis were masker noise (white vs. cubed vs. interrupted), subject, and S/N ratio. Masker noise and subject were treated as factors (categorical variables) whereas S/N ratio was treated as a continuous variable. Using the GLIM 3.77 software (Payne [22]), a number of hierarchically-nested models (based on the logit link function) were fitted. The standard logit link function was modified (via "Abbott's formula" [7]) to take account of the fact that even at extremely low S/N ratios, subject performance should not fall to 0% correct, but rather to about 8.3% (1/12), because there are only 12 possible responses. In practice, because most of the points on the performance intensity functions were well above this level of performance, the correction had no influence on the major aspects of the statistical findings.

Changes in deviance (related to the adequacy of the fit of the model) were used to assess the statistical significance of excluding various terms of the saturated model (which included all the interactions and main effects of the three predictors). All statistical tests used a 0.05 significance level. The final model, incorporating three main effects (collapsing the white and cubed noise conditions into one class) plus a second-order interaction term between the slope of the performance intensity function and subject, fitted the data extremely well ($p \approx 0.9$, based on the generalized Pearson's χ^2). This model allowed the following conclusions:

1) There were no differences between the results obtained with white and cubed noise, either in speech reception threshold or slope.

2) Interrupted noise led to a lower speech reception threshold than the other two maskers, but its performance intensity function did not differ in slope from the other two conditions.

3) Subjects differed from one another in both the slope and intercept of the performance intensity function, but there were no interactions with masker noise type. In other words, each subject performed identically with white and cubed noise, and had a lower speech reception threshold with interrupted noise.

Why, then, does cubed 7 noise behave similarly to white noise as a speech masker, despite the severe (instantaneous) fluctuations of its power? One important factor arises from a consideration of the properties of peripheral auditory filtering. In particular, auditory filters have a temporal response which will "smooth" out any rapid, instantaneous fluctuations, except, possibly, in the higher frequency channels, where the temporal responses of the filters are faster. Samples of the white and cubed noise used in this experiment were passed through a digital simulation of three auditory filters (so-called gammatone filters with centre frequencies at 250 Hz, 1000 Hz and 5000 Hz – Patterson and Cutler [21]). The resulting waveforms are shown in Fig. 3. Instantaneous fluctuation can be measured using the fourth moment of the waveform and such an analysis was performed on the output of these filters. Only in the channel with the highest centre frequency of 5 kHz could differences be

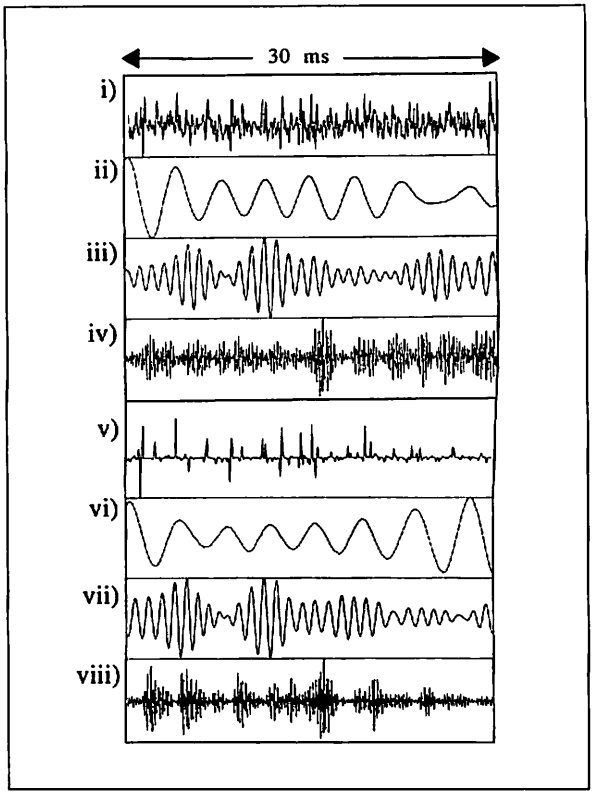

Fig. 3. The waveforms that result from passing white noise (top left) and cubed noise (top right) through digitally simulated auditory filters at three centre frequencies. The filtered waveforms are shown underneath the noise samples from which they are derived. i) White noise and the same noise after filtering at ii) 250 Hz, iii) 1 kHz and iv) 5 kHz; v) cubed noise and the same noise after filtering at vi) 250 Hz, vii) 1 kHz and viii) 5 kHz.

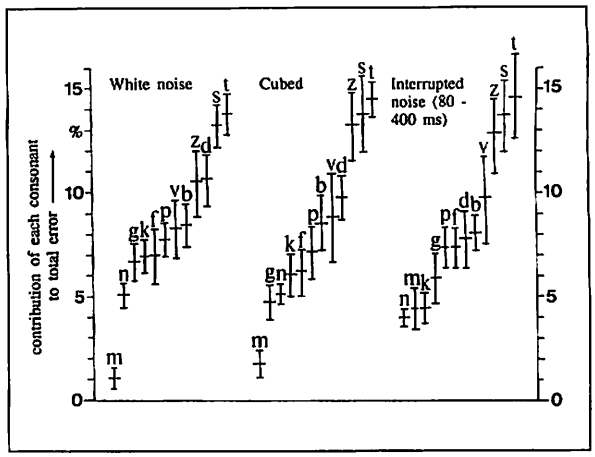

Fig. 4. The relative difficulty of identification associated with each consonant in each of the three noise environments studied in experiment 1. Using results from the second part of the experiment, the numbers of incorrect perceptions of each consonant as a percentage of the total number of errors were calculated for each subject. The diagram shows the results of this calculation averaged over the six subjects.

detected between the fluctuations caused by cubed noise and ordinary white noise. Such differences are unlikely to influence speech perception in this experiment (especially since all material was played out through a 5.6 kHz lowpass filter).

Rudimentary phonetic analyses of the results did not reveal any striking differences between the types of confusions made by subjects listening against the three noise backgrounds for either part of the experiment. The results from measurements of the performance intensity functions were used to calculate the percentage of the total number of errors made by each subject, for each noise type, that was attributable to incorrect responses to each speech token. These statistics were averaged over the set of subjects to produce Fig. 4.

This diagram gives an indication of the relative difficulty associated with each of the speech sounds in each noise environment. For each noise, the speech sounds follow in order of apparent difficulty. The relative intelligibility of the speech tokens does not appear to be greatly influenced by the types of speech masker used. Neither does the general ordering observed here conflict greatly with, for example, the extensive study of Miller and Nicely [20]. Exceptionally, our subjects had apparently little difficulty in perceiving /g/, which was reported by Miller and Nicely [20] as being one of the least intelligible consonants. However, the use, in our experiments, of only one token of each of a smaller set of consonants makes it difficult to compare results with those of Miller and Nicely [20]. Also, for speech at the widest bandwidth used, and S/N ratios of 0, −6 and −12 dB, Miller and Nicely found the consonant /ʒ/ to be most often confused with /g/ and /d/, and the omission of /ʒ/ from our test set would undoubtedly have helped subjects to recognise /g/ correctly.

3. Experiment 2

It had been thought that the interrupted noise might cause subjects to make confusions that were more disparate with respect to their phonetic features. An increase in the randomness of the subject's decision-making might also have led to a flattening of the performance intensity function. Since neither of these effects were observed, it was thought worthwhile to run a second experiment, using noise which fluctuated on an even longer time-scale.

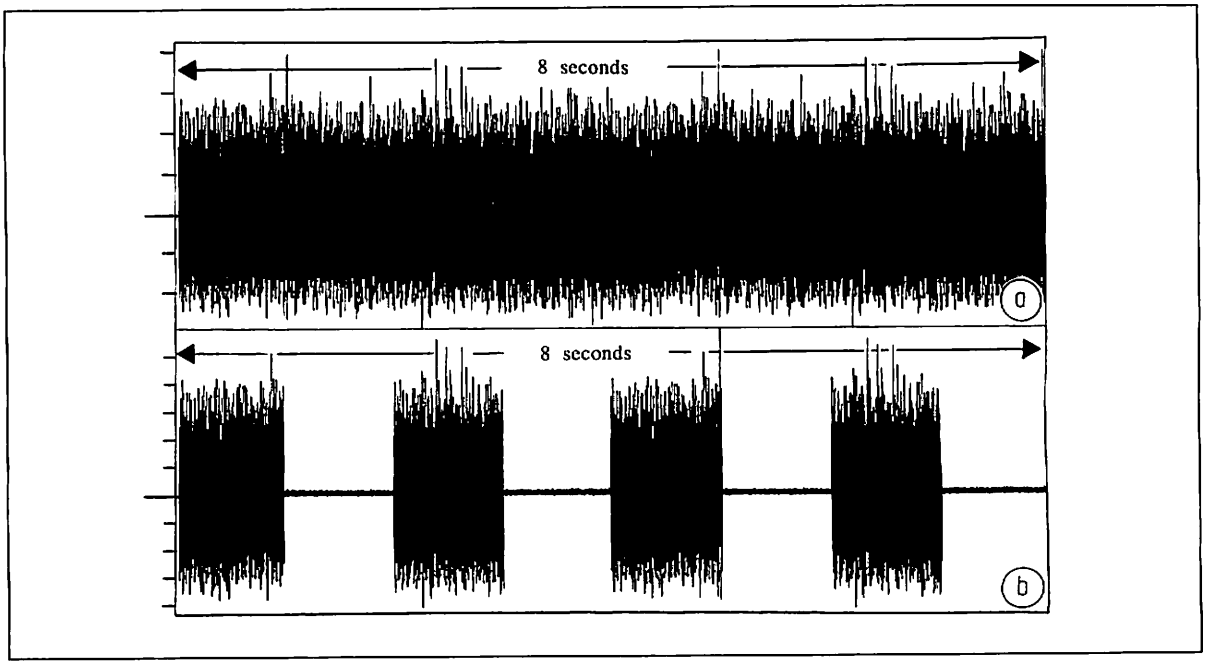

Fig. 5. Noise samples used in experiment 2: a) white noise and b) the same noise with regularly spaced interruptions of one second duration, every other second.

3.1. Method

3.1.1. Subjects

The subjects were in 2 groups, one for each part of the experiment, consisting of 6 and 4 volunteers, respectively.

3.1.2. Stimuli

The speech material was identical to that used in the first experiment. Two samples of masking noise were compared here. The first was an 8 second sample of Gaussian white noise (Fig. 5a) sampled at 12.8 kHz. The second was created from the first by attenuating every other one second section by 30 dB (Fig. 5b).

3.2. Procedure

As in experiment 1, the speech reception threshold was determined using adaptive procedures and performances intensity functions were obtained using the same experimental protocol as before. This time, however, the performance intensity functions were determined using 4 subjects listening to test material at $+12$, $+6$, $+3$, 0, -3, -6, -12 dB relative to a previously estimated speech reception threshold.

3.3. Results and discussion

The mean speech reception thresholds for the two noises and two procedures, averaged over the subjects, are given in Table II. The mean performance intensity functions for the two noises are shown in Fig. 6. Adaptive testing showed a release of masking with the interrupted noise ($p < 0.0001$, on the basis of an ANOVA).

As for experiment 1, the effects of various factors on the performance intensity functions were evaluated using a generalised linear model incorporating masker noise type (white vs. interrupted), subject and S/N ratio. Unlike the analysis of experiment 1, even the highest order interaction term was marginally significant ($p \approx 0.04$) indicating that the slopes of the performance intensity functions varied with masker type

Table II. Speech reception thresholds (SRTs) measured for two different masking noises using two different techniques. Adaptive SRTs are the mean of two tests from each of 6 subjects, while performance intensity (PI) estimates were obtained by taking the mean SRT for each of four subjects determined from a logistic fit to the individual PI functions. SRTs are indicated on an arbitrary scale in which the SRT for continuous white noise of experiment 1 was set to 0.0 dB.

Noise type	Adaptive		PI	
	SRT dB	Standard error dB	SRT dB	Standard error dB
Continuous white	+2.40	±0.78	+1.45	±1.26
Interrupted (1 s duration)	−6.34	±0.85	−14.94	±1.61

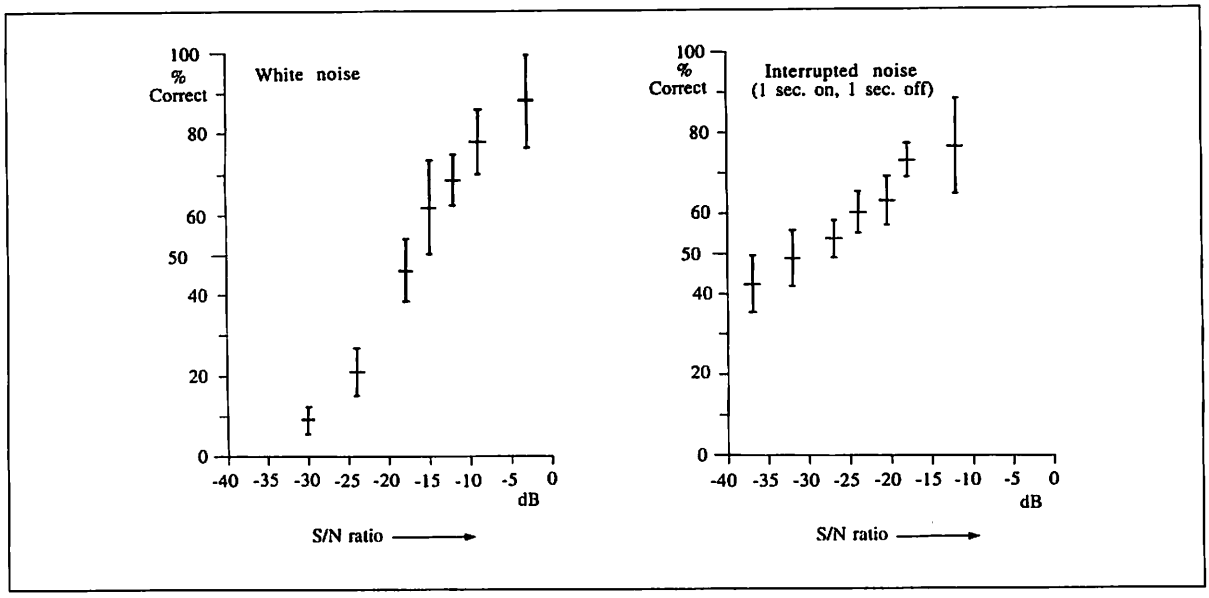

Fig. 6. The performance intensity functions for white noise and interrupted noise (1 s on 1 s off), derived by combining results for four subjects listening to the test material over a range of S/N levels.

and subject. By excluding this term, however, it was possible to construct a much simpler model which, although not strictly adequate, still fitted the main trends in the data reasonably well. This reduced model led to a significant generalized Pearson's χ^2 value ($p \approx 0.012$), indicating that the inadequacies of the model were unlikely to be solely due to chance. A comparison of the fitted values of the saturated model (including all the factors and their interactions, and which fit the data well) to those from the reduced model showed that the worsening of the fit was generally spread across the entire data set, and not confined to small regions of the parameter space. It is also interesting to note that a standard logistic procedure, which does not assume a lower asymptote of 8.3%, led to the same conclusions as the model which included this lower asymptote, but fitted the data better ($p \approx 0.06$ on the basis of generalized Pearson's χ^2 for the reduced model). In summary, although the lack of fit in the model needs further investigation, our reduced model appears to be adequate for the purposes used here.

This model, in which only the three main effects, plus a single second-order interaction between noise type and S/N ratio were necessary, allowed the following conclusions:

1) For a particular masker, there were no significant differences among the slopes of the performance intensity functions exhibited by different subjects, but there were differences in speech reception threshold.

2) Both the intercept and the slope of the performance intensity function were significantly different for the two maskers ($p < 0.005$), with the performance intensity function being flatter for the interrupted noise. A t-test on the speech reception thresholds obtained by logistic fits of the individual performance intensity functions, showed the speech reception threshold for interrupted noise to be significantly lower than that obtained for white noise ($p < 0.005$).

Therefore, as in experiment 1, interrupted noise led to a release of masking for both testing techniques. The "glimpsing" effect, referred to above, can still be used to explain the general release of masking observed for this interrupted noise. Here, however, it may be caused by a consonant still being intelligible even when its beginning or end falls within the louder burst, the unmasked part of it proving sufficient for identification.

By comparing Tables I and II, it can be seen that the speech reception threshold for white noise was slightly lower in experiment 1 than in experiment 2. It may be that the longer sample of noise used in experiment 2 increases the difficulty of the task by adding uncertainty as to when the token will be presented.

Table II also shows a large difference in experiment 2 between the speech reception threshold for interrupted noise measured adaptively and from the performance intensity function – larger than would appear to be attributable to practice or group differences. In our estimation, this results from a deterioration in the efficiency of the adaptive procedure because performance is less sensitive to changes in S/N level for interrupted noise (reflected in the flattening of the performance intensity function). Therefore, the

adaptively measured value of speech reception threshold for the fluctuating noise is misleading. If the adaptive test had been extended, it is likely that the average presentation level would have settled at a lower value.

The nature of the fluctuating noise used in this experiment is such that for a range of overall S/N levels, half of the speech can be almost completely masked whilst that half of the speech occurring between the louder bursts remains reasonably intelligible. The incorrect decisions made by the observer are, therefore, very dependent on when the speech token occurs in relation to the bursts of noise. Thus, the types of phonetic confusions observed for these conditions will be more random, since the acoustic cues relating to the phonetic features of the speech have less bearing on their intelligibility. A scatter-plot of the number of voicing errors against the number of incorrect responses is shown in Fig. 7. This diagram illustrates how the normally robust feature of voicing can be less easily detected in noise that is interrupted on long time-scales than in white noise, when compared at S/N levels that lead to similar intelligibility.

Fig. 8 shows the relative difficulty of identification (derived by the means described in section 2.3) associated with each speech sound occurring in each of the noises. The bunching observed in the right hand part of the figure reflects how subjects' performances appear to be less-related to the type of speech sound presented.

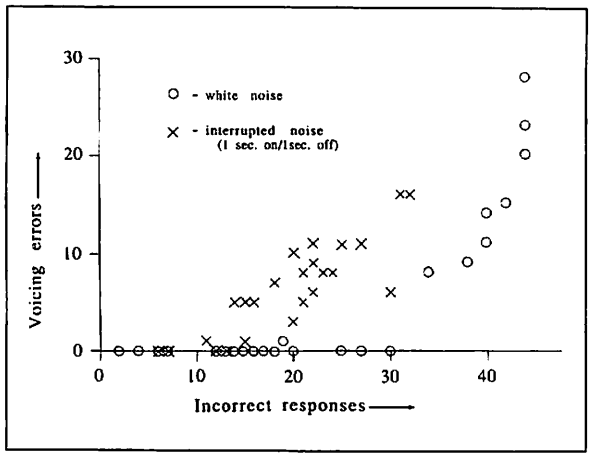

Fig. 7. Scatter-plot of the number of voicing errors against the total number of incorrect responses for each testing session used for determination of the performance intensity function in experiment 2. The number of points associated with the white noise positioned along the horizontal axis implies that even when a large number (up to 30 items) is incorrectly perceived, there may be no voicing errors. However, with the interrupted noise, perception of this robust acoustic cue is clearly disrupted, since there is a general clustering of points closer to the diagonal, where voicing errors increase proportionally to incorrect responses.

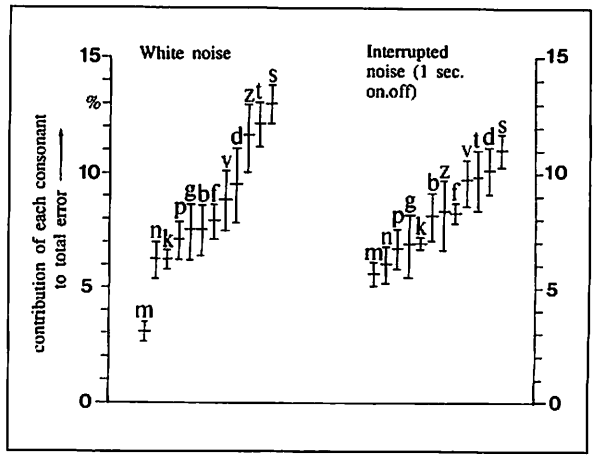

Fig. 8. The relative difficulty of identification associated with each consonant in the two noise environments studied in experiment 2. Using results from the second part of the experiment, the numbers of incorrect perceptions of each consonant as a percentage of the total number of errors were calculated for each subject. The diagram shows the results of this calculation averaged over the four subjects.

4. Experiment 3

Having made these preliminary investigations as to how perception at the phonetic level may be influenced by masker fluctuation, it was decided that the variation of intelligibility of this type of speech corpus should be measured with respect to the duration of the quieter section of the fluctuation.

4.1. Method

4.1.1. Subjects

The subjects consisted of 6 volunteers.

4.1.2. Stimuli

There had previously been some concern that using only one token of each speech sound might give results which were unrepresentative. To ensure against this, the speech material used in these tests was enlarged to four examples each of the twelve intervocalic consonants used previously, spoken by the same female speaker. To improve the amount of high frequency information being included, the data was sampled at 44.1 kHz, and played out through a lowpass filter set at 10 kHz.

Seven types of noise signal were used in this experiment. The first noise signal was 10 seconds of white noise. The other six were created by interrupting copies of this signal at six different rates (100, 75, 50, 25, 10 and 5 Hz; noise-time fraction = 0.5). These noise signals can be considered to represent a range of

conditions with respect to fluctuation duration (5, 6.7, 10, 20, 50 and 100 ms), the magnitude of fluctuation being held constant ($-\infty$ dB, expressed as the level of the quiet section relative to the long-term average).

4.2. Procedure

The speech reception threshold was measured using the adaptive procedure outlined above, except that settling times were improved by increasing the step size to 1 dB. In an attempt to minimize the effects of order, three subjects were tested with the 7 noises in increasing order of fluctuation duration, and three subjects in reverse order.

In experiments 1 and 2, the whole noise signal was presented, with the speech signal starting at a random point in it. In experiments 3 and 4, a random portion of the noise signal, of duration equal to that of the speech sound, was chosen before each stimulus presentation, for mixing with the speech signal.

4.3. Results and discussion

Speech reception thresholds for each noise condition were determined, relative to that for the white noise condition, by averaging over the six subjects. The results in Fig. 9 clearly show a decrease in masking effectiveness with increasing duration of fluctuation ($p < 0.0001$ by an ANOVA). This concurs reasonably well with the results of Miller and Licklider [19] for monosyllables, who found that maximum speech scores were obtained when the durations associated with fluctuation were in the region of 50 ms.

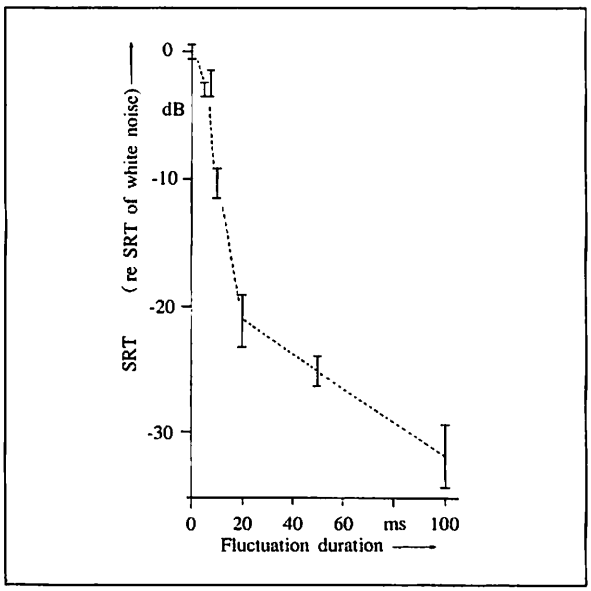

Fig. 9. The speech reception threshold for consonants in simple on/off fluctuations as a function of the duration of the fluctuation.

5. Experiment 4

The fluctuation of the interrupted broad-band maskers in the previous experiment and those used by Miller and Licklider [19] were created by effectively switching a white noise source on and off at different rates. They can be characterized simply in terms of a noise-time fraction (the fraction of time the noise was on) and the rate of interruption, since the power of the noise during the off-periods was always attenuated to zero. If the power had been attenuated by only a few dB (relative to the overall average) one may also assume that the masking release would have been less, so the magnitude of the fluctuation must be considered important. Experiment 4 investigates how the magnitude of the masker fluctuation influences speech intelligibility.

5.1. Method

5.1.1. Subjects

The subjects consisted of 6 volunteers.

5.1.2. Stimuli

For this experiment, 8 different noise signals were generated. The first was 10 seconds of white noise. The other seven were generated from copies of this signal by attenuating every other 50 ms section by a constant amount (-6, -12, -18, -24, -30, -36 and $-\infty$ dB). Thus, a set of fluctuating noises were created, each containing 50 ms fluctuations, but representing a range of conditions with respect to fluctuation magnitude. It is often more convenient to measure the magnitude of fluctuations relative to the long-term average power of the noise (see below). On such a scale, the magnitudes of the fluctuations in this experiment were 0, -4.0, -9.3, -15.1, -21.0, -27.0, -33.0 and $-\infty$ dB, respectively. All noise conditions exhibited, of course, flat frequency spectra.

The speech material was the same as that used for experiment 3.

5.2. Procedure

As in experiment 3, the tests were arranged such that 3 subjects heard the noise conditions in order of increasing magnitude of fluctuation, and 3 in decreasing order.

5.3. Results and discussion

Speech reception thresholds for each noise condition were determined, relative to that for the white noise condition, by averaging over the six subjects. These

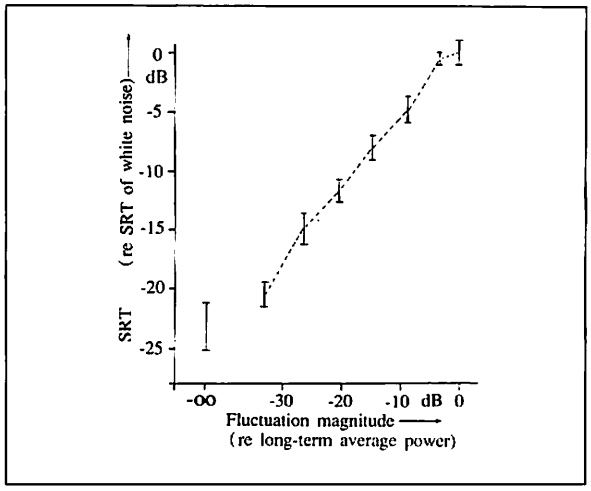

Fig. 10. The speech reception threshold for consonants as a function of the magnitude of 50 ms fluctuations in white noise. The magnitude of the fluctuation has been plotted as the level of the quiet section relative to the long term average.

have been plotted in Fig. 10 against the magnitude of fluctuation.

As expected, there is a large difference in masking effectiveness between the white and on/off noise conditions (about 23 dB). Also, the speech reception threshold decreases in a linear fashion with increasing magnitude of fluctuation, at a rate of about 0.64 dB per dB over a 33 dB range (the correlation coefficient being 0.994). Fluctuation magnitudes of the order of 10 to 20 dB are quite unlikely in environments such as traffic noise or speech babble, so the largest effects of glimpsing are probably restricted to special – but not uncommon – instances, such as competing speech from a single talker.

6. Quantifying the fluctuation of a masker

6.1. Summary of the effects of masker fluctuation on speech perception

Rapid fluctuations of short duration have no influence upon masking effectiveness and noise backgrounds characterized by such fluctuations may be considered equivalent to continous noises of similar frequency content. Such rapid fluctuations are effectively removed by the limited temporal resolution of the auditory system.

When the duration of the fluctuation extends beyond about 10 ms, the phenomenon of "glimpsing" begins to provide a significant release from masking. Relative intelligibility, at least at the segmental level, does not appear affected by the perceptual processes of glimpsing, and the slope of the performance intensity function remains similar to that for continuous noise of the same spectral content.

Glimpsing increases with duration until, for fluctuations of long duration, a particular speech sound is either completely masked or completely intelligible. Correspondingly, the performance intensity function flattens and the pattern of phonetic confusions becomes randomised.

Glimpsing is sensitive to the level of the quiet sections during which it occurs, and the release of masking appears to be approximately proportional to the magnitude of the fluctuation for fluctuations down to about – 30 dB (relative to average power).

6.2. Definition of a fluctuation

Having outlined those characteristics of fluctuations which appear critical to their effect upon speech perception, it may now be appropriate to define the power fluctuation of a masker in terms of these dimensions:

A fluctuation is the excursion of the power envelope below a certain value (relative to the long-term average), characterized by the magnitude and direction (±) of the excursion from that value and its duration.

Fig. 11 shows a schematic power envelope (here defined as a time-averaged version of the square of the original waveform) containing a single idealised fluctuation. This can be described as an excursion of the power envelope below – 8 dB (relative to the long-term average) for 120 ms. Note that it is the shape of the power envelope that is of concern to us, since it is the size of the larger "gaps" in the envelope that determine the probable masking release. A frequency anal-

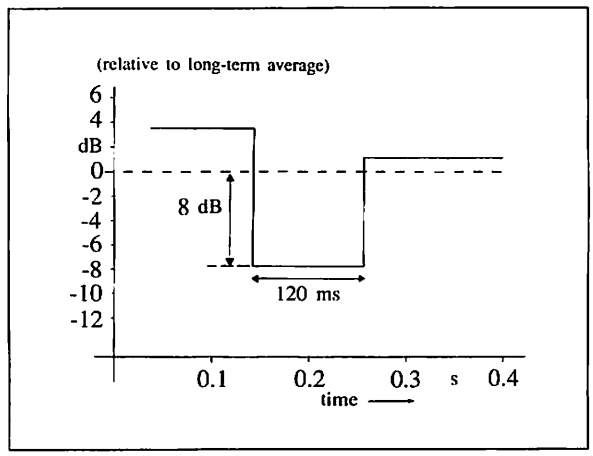

Fig. 11. An idealised fluctuation in power. The diagram shows the power of a signal dropping below a certain reference threshold for a certain length of time. If the power threshold is defined relative to the long-term average, this fluctuation can be described as having a magnitude of – 8 dB (relative to the average) and a duration of 120 ms.

ysis of the power envelope (see, for example, van Dijk and Wit [3]) will not, therefore, yield the appropriate information. For example, the power envelope of a randomly interrupted signal is a random series of pulses. A frequency analysis of such a signal will reveal high-frequency components which are not derived from interruptions occurring at a fast rate, but which bear only a complex relationship to the lower "frequency" interruptions we are trying to quantify. It would be difficult to interpret the results of such an analysis in terms of "gaps" in the masker, without involving a complex analysis of the phase components which hold vital temporal information.

The fluctuations investigated in the experiments described above consist of energy changes in various frequency regions which are time-correlated – that is, they are comodulated. Although real noise environments cannot be expected to behave so uniformly, it seems reasonable, as a first approximation, to restrict the concept of masker fluctuation to this simple comodulated case. Thus, we consider our "real" noise masker as a signal of constant spectral content over its total duration but whose rms, measured over shorter time-periods, is changing. Recent investigations have indeed indicated that the kind of uncomodulated noises that lead to significant releases from masking may be quite unlikely to occur in "real" environments (Festen and Plomp [6]; Howard-Jones and Rosen [13, 14]).

6.3. A quantitative fluctuation analysis

Given the assumptions and simplifications made above, it is now possible to specify a fluctuation analysis of any masker based on the three dimensions of magnitude, duration and frequency of occurrence, without generating an excess of data for interpretation or display. The analysis consists of two parts: the generation of a smoothed power envelope and the application of an algorithm to measure and count the fluctuations.

6.4. Generation of the power envelope

The generation of the power envelope proceeds by first squaring the waveform of the noise signal and then time-averaging it with a raised-cosine window.

Two types of acoustic feature may be thought to determine the shape of the original waveform, and the shape of the signal derived by squaring it. The first type is related to the frequency content of the signal, since any waveform is required to fluctuate in order to carry the acoustic information we perceive as frequency information. (Even the power of a continuous sinusoid of frequency F Hz, may be thought to fluctuate at a rate of $2F$ Hz.)

The power fluctuations associated with the concept of a power envelope, however, are of the type which can be heard out in the original waveform as changes in signal level. This second type of acoustic information is derived from features in the squared waveform that tend to consist of lower frequency components than those related to waveform frequency information, and can be extracted by using a time-averaging window to filter out the higher frequency components. Unfortunately, some degree of overlap exists between the frequency ranges of these two categories of fluctuation. However, we are helped in this instance by the fact that we are interested only in "gaps" long enough to "glimpse" speech information, and not with all fluctuations in the original waveform which are simply perceivable. A raised-cosine time-window (with a duration between null points of 14 ms) was found, by informal experimentation, to be appropriate for smoothing out the microstructure of the power envelope, without removing the shorter fluctuations associated with masking effectiveness. The decision to use a raised cosine window was based on empirical considerations, but future attempts to develop a more appropriate temporal window may benefit from applying knowledge about the human auditory system. A simple consideration of forward and backward masking characteristics has already been used to link the glimpsing performance of hearing-impaired listeners to deficits in temporal resolution (Festen [5]).

6.5. An algorithm to count fluctuations

After smoothing, the duration of each excursion of the power envelope below each of a set of pre-defined power thresholds (dB relative to the average) is measured. The value of each duration is discretized by determining which amongst a set of duration ranges it falls within. The total number of fluctuations of each type (characterised by magnitude and duration ranges) is recorded. The distribution is then calculated in terms of the estimated total percentage of all time that the power envelope falls below each power threshold for each range of duration.

6.6. Example of analysis results

Fig. 12 shows an example of the fluctuation analysis applied to a 4-s white noise signal interrupted at a rate of 6.25 Hz (noise-time fraction = 0.5, resulting in fluctuations of 80 ms). Fig. 12a) shows the original waveform, Fig. 12b) the waveform after squaring and Fig. 12c) after time-averaging with the exponential window. The distribution of the fluctuations, created by the procedure outlined above, is shown in Fig. 12d). The length of each box (corresponding to a

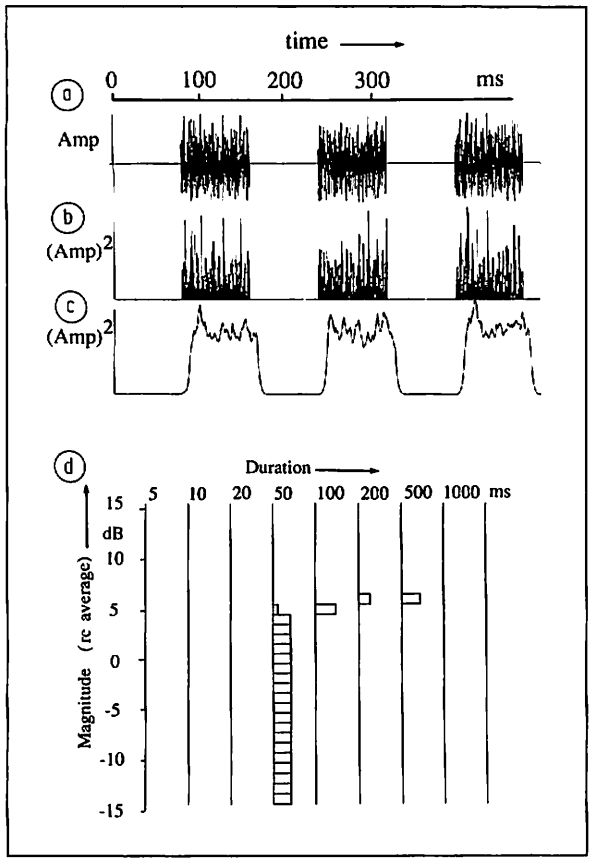

Fig. 12. A proposed fluctuation analysis: a) Waveform of an interrupted white noise and b) the same waveform after squaring and c) the power envelope of the signal created by averaging the previous signal with a time window. Finally, the distribution of fluctuations in this power envelope are displayed in d), where the size of each box refers to the proportion of total time that the signal was below a certain power threshold (given by the vertical axis, relative to the long-term average) for various duration ranges (given by the horizontal axis). The distance between any two vertical lines corresponds to 100% of the total signal duration.

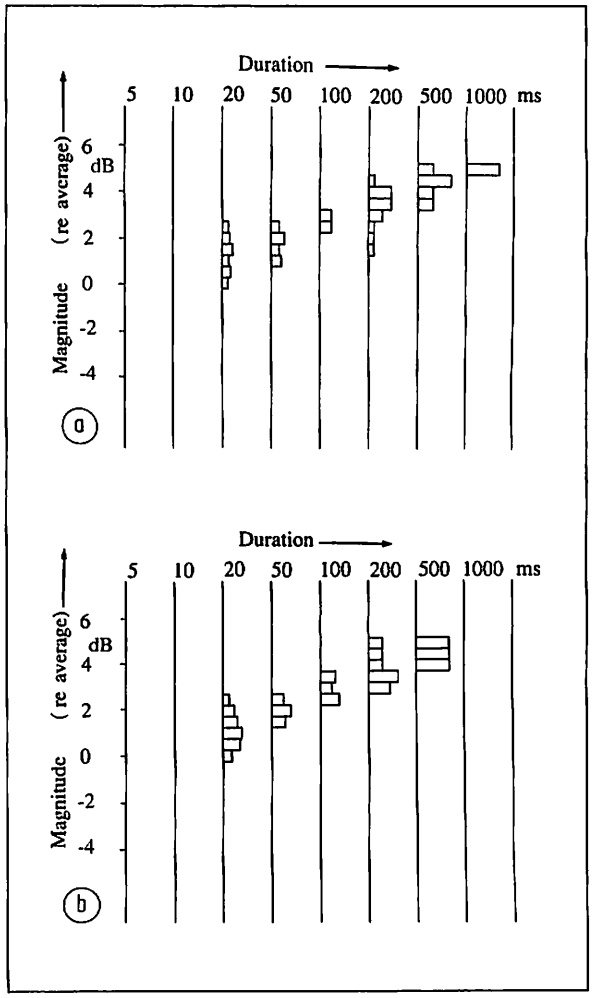

Fig. 13. Results of the proposed fluctuation analysis for a) cafeteria noise and b) shaped white noise of the same long-term frequency content. The distribution of fluctuations appears similar in both cases, and no significant difference in the masking properties of the two noises could be detected by experiment.

particular power value and duration range) is proportional to the total time that the power envelope has fallen below that power value for durations within that range. Further results from this type of analysis are shown in Fig. 13 for a) cafeteria noise and b) shaped white noise.

Several features characterising the fluctuation of a noise background are revealed by such an analysis:

i) On/off noises characterized by regular interruptions appear as a column of boxes reaching down to large negative values of magnitude (see Fig. 12d).

ii) Fluctuations of lower magnitude (e.g. interruptions during which the off-period was equivalent to only partial attenuation) reveal themselves as shorter columns which do not extend so far in the negative magnitude direction. Even continuous shaped white noise contains fluctuations according to our definition, but these are very restricted in their magnitudes (Fig. 13b).

iii) The probable masking release caused by such interruptions will depend, not only upon the "depth" of the column, but on its position along the horizontal (duration) axis. For example, fluctuations in noise power belonging to the 50–100 ms range will be more likely to contribute to masking release than fluctuations of similar magnitude and prevalence occurring in lower ranges, especially those below 10 ms.

Obviously, further empirical results concerning the effects of masker fluctuation on speech perception are needed before this type of analysis can be calibrated to provide a quantitative prediction of masking effectiveness. As a final example, however, Fig. 13 shows the

results of the analysis for a) cafeteria noise and b) a white noise shaped to possess the same long-term frequency characteristics. The fluctuation characteristics also appear similar for both noises. The two signals might thus be expected to possess similar masking characteristics. In an adaptive test using six subjects (similar to that described for experiment 1 and 2) no significant difference was found for the speech reception threshold associated with the two noises. There may, therefore, be no practical reason for carrying out speech perception tests using cafeteria noise in preference to shaped white noise, even though cafeteria noise may appear more "realistic".

7. Conclusions

When considering their effect upon speech perception, masker fluctuations may be divided into three categories of duration.

Short-term fluctuations (of duration less than about 5 ms) have no effect on speech perception. Such fluctuations are effectively removed by the temporal smearing of the auditory system.

Fluctuations of medium duration (10–200 ms, say) bring about an observable release of masking through glimpsing of several parts of individual acoustic cues, but are not likely to produce a pattern of phonetic confusions any different to that observed with steady-state noise of the same spectral shape. The performance intensity function is shifted horizontally by the masking release but its shape remains the same.

Fluctuations of longer duration also give rise to masking release, through the listener's ability to glimpse the less obscured beginning or end of the speech sound. Here, however, the stronger dependence of recognition upon the chance position in time of the speech sound relative to the fluctuations, causes a flattening of the performance intensity functions and a randomization of the pattern of phonetic confusions.

Masking release, through glimpsing, also depends strongly upon the magnitude of the fluctuation. With the speech material used above, it was found that the speech reception threshold was reduced by 0.63 dB with every dB change in fluctuation magnitude over a 30 dB range (measured as the excursion of the power envelope relative to the long-term average).

A method has been proposed for measuring masker fluctuation in terms of those dimensions considered critical to speech perception. It is hoped that such an approach may provide a useful tool for indicating whether the fluctuation of a particular environmental noise prevents it from being characterized simply in terms of its power spectrum. Where such signals cannot be treated as continuous, this type of measurement may serve as a basis for a future predictor of intelligibility.

Acknowledgements

First thanks must go to: Frederika Holmes for many hours spent in translating two papers from German into English; J. Sotscheck for providing a technical report and helpful comments on hte manuscript; and Christian Benoît for translating the English abstract into French. This research could not have been done without the support of the Department of Phonetics & Linguistics, University College London. Stuart Rosen was supported primarily by the Medical Research Council (UK), with timely bridging funds from the Hearing Research Trust and the Heinz & Anna Kroch Foundation.

References

[1] Bosman, A. J., Speech perception by the hearing impaired. Ph.D. thesis, University of Utrecht, Netherlands 1989.
[2] Cooper, J. C., Cutts, B. P., Speech discrimination in noise. J. Sp. and Hear. Res. **14** [1971], 332.
[3] van Dijk, P., Wit, H. P., Amplitude and frequency fluctuations of spontaneous otoacoustic emissions. J. Acoust. Soc. Amer. **88** [1990], 1779.
[4] Dirks, D. D., Morgan, D. E., Dubno, J. R., A procedure for quantifying the effects of noise on speech recognition. J. Sp. Hear. Disorders **47** [1982], 114.
[5] Festen, J. M., Speech-reception threshold in fluctuating background sound and its possible relation to temporal auditory resolution. In: The Psychophysics of Speech Perception, edited by M. E. H. Schouten (NATO ASI Series D), Nijhoff (The Netherlands), 1987.
[6] Festen, J. M., Plomp, R., Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. J. Acoust. Soc. Amer. **88** [1990], 1725.
[7] Finney, D. J., Probit analysis. 3rd ed. Cambridge University Press, Cambridge [1971].
[8] Foster, J. R., Haggard, M. P., The four alternative auditory feature test (FAAF)-linguistic and psychometric properties of the material with normative data in noise. Br. J. Audiol. **21** [1987], 165.
[9] French, N. R., Steinberg, J. C., Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Amer. **19** [1947], 90.
[10] Gordon-Salant, S., Phoneme feature perception in noise by normally-hearing and hearing-impaired subjects. J. Sp. Hear. Res. **28** [1985], 87.
[11] Hartmann, W. M., Pumplin, J., Noise power fluctuations and the masking of sine signals. J. Acoust. Soc. Amer. **83** [1988], 2277.
[12] Hawkins, J. E., Stevens, S. S., The masking of pure tones and speech by white noise. J. Acoust. Soc. Amer. **22** [1950], 6.

[13] Howard-Jones, P. A., Rosen, S., Speech perception in "chequerboard" noise. Abstract, Br. J. Audiol. **26** [1992], 187.
[14] Howard-Jones, P. A., Rosen, S., Uncomodulated glimpsing in "checkerboard" noise. J. Acoust. Soc. Amer. 1993, in press.
[15] Kalikow, D. N., Stevens, K. N., Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. J. Acoust. Soc. Amer. **61** [1977], 1337.
[16] Keith, R. W., Talis, H. P., The effects of white noise on normally hearing and hearing impaired subjects. J. Sp. Hear. Res. **28** [1972], 87.
[17] Levitt, H., Transformed up-down methods in psychoacoustics. J. Acoust. Soc. Amer. **41** [1971], 467.
[18] Miller, G. A., The masking of speech, Psychol. Bull **44** [1947], 105.
[19] Miller, G. A., Lickider, J. C. R., The intelligibility of interrupted speech. J. Acoust. Soc. Amer. **22** [1950], 167.
[20] Miller, G. A., Nicely, P. E., An analysis of some perceptual confusions among some English consonants. J. Acoust. Soc. Amer. **27** [1955], 338.
[21] Patterson, R. D., Cutler, A., Auditory preprocessing and recognition of speech. In: Cognitive Psychology – Research Directions in Cognitive Science. Baddely, A., Bernsen, N. O., Eds., Adacemic Press, London 1989.
[22] Payne, C. D., The GLIM system release 3.77 manual – Edition 2. Numerical Algorithms Group Ltd., Oxford, UK 1985.
[23] Sotscheck, J., Messungen zur Sprachverständlichkeit bei additiv wirkenden Störsignalen. Teil 1: Dokumentation der Versuchsbedingungen. Techn. Ber. Forsch.-Inst. FTZ, Dtsch. Bundespost, 13 TBr 14, Dez. 1982.
[24] Sotscheck, J., Sprachverständlichkeit bei additiven Störungen. Acustica **57** [1985], 257.
[25] Speaks, C., Karmen, J. L., Benitez, L., Effect of a competing message on synthetic sentence identification. J. Sp. Hear. Res. **10** [1967], 390.