Listening to speech in a background of other talkers: Effects of talker number and noise vocoding

Stuart Rosen[a)]
UCL Speech, Hearing and Phonetic Sciences, 2 Wakefield Street, London WC1N 1PF, United Kingdom

Pamela Souza
Department of Communication Sciences and Disorders, Knowles Hearing Center, Northwestern University, 2240 Campus Drive, Evanston, Illinois 60208

Caroline Ekelund[b)]
UCL Speech, Hearing and Phonetic Sciences, 2 Wakefield Street, London WC1N 1PF, United Kingdom

Arooj A Majeed
UCL Ear Institute, 332 Grays Inn Road, London WC1X 8EE, United Kingdom

(Received 29 February 2012; revised 5 February 2013; accepted 11 February 2013)

Some of the most common interfering background sounds a listener experiences are the sounds of other talkers. In Experiment 1, recognition for natural Institute of Electrical and Electronics Engineers (IEEE) sentences was measured in normal-hearing adults at two fixed signal-to-noise ratios (SNRs) in 16 backgrounds with the same long-term spectrum: unprocessed speech babble (1, 2, 4, 8, and 16 talkers), noise-vocoded versions of the babbles (12 channels), noise modulated with the wide-band envelope of the speech babbles, and unmodulated noise. All talkers were adult males. For a given number of talkers, natural speech was always the most effective masker. The greatest changes in performance occurred as the number of talkers in the maskers increased from 1 to 2 or 4, with small changes thereafter. In Experiment 2, the same targets and maskers (1, 2, and 16 talkers) were used to measure speech reception thresholds (SRTs) adaptively. Periodicity in the target was also manipulated by noise-vocoding, which led to considerably higher SRTs. The greatest masking effect always occurred for the masker type most similar to the target, while the effects of the number of talkers were generally small. Implications are drawn with reference to glimpsing, informational vs energetic masking, overall SNR, and aspects of periodicity.
 [http://dx.doi.org/10.1121/1.4794379]

PACS number(s): 43.72.Dv, 43.71.Es, 43.71.Rt, 43.66.Ts [MAA] Pages: 2431–2443

I. INTRODUCTION

As Miller so elegantly wrote in his pioneering paper of 1947, "It has been said that the best place to hide a leaf is in the forest, and presumably the best place to hide a voice is among other voices" (Miller, 1947, p. 118). It should come as no surprise then that trying to understand 1 talker in a background of others is among the most difficult communication situations. The listener must attend to an acoustic signal from a specific talker among other background signals which will be similar to the target in spectral and temporal features, may come from a similar location, may be less or more intense, and may also contain similar semantic content.

A crucial determinant of the masking effectiveness of other speech is the number of talkers that are present in the background, especially for small numbers of talkers. Not only is this manipulation one which is ecologically valid, it can also serve as a foundation for testing different ideas about what factors make a masker more or less effective. Miller (1947) noted that a single talker was not a particularly potent masker, reporting that for a fixed SNR performance for identifying isolated words decreased as the number of talkers in the babble was increased over 1, 2, 4, and 6 talkers. There was also evidence of increasing performance as the number of talkers in the masker went from 6 to 8, the maximum tested. Freyman *et al*. (2001) used nonsense sentences as targets, and found performance decreased sharply with 2 talkers in the masker as compared to 1 talker. In a separate but similar study using babble maskers consisting of 2–10 talkers, performance was worse for 2 talkers and generally improved as more talkers were added (Freyman *et al*., 2004). Simpson and Cooke (2005) used vowel-consonant-vowel syllables (VCVs) as target material and found the worst scores for 8 talkers, with clear increases in performance as the number of talkers in the babble was increased to 512. The main goal of the present study was to establish how the intelligibility of meaningful sentences in a competing babble of other talkers changes as the number of talkers in the babble is varied. At the same time, we also tried to establish what factors were responsible for these trends.

What kind of processes might be operating that would explain this non-monotonic change in performance? Brungart (2001) delineated two major ways in which background sounds may interfere with perception of the target

[a)]Author to whom correspondence should be addressed. Electronic mail: stuart@phon.ucl.ac.uk
[b)]Present address: Audiology Department, Royal Berkshire Hospital, Reading, England.

signal—*energetic* and *informational* masking. Energetic masking (EM) is posited to be directly related to the presence of masker energy in the same frequency region(s) as energy in the target signal, causing reduced audibility at a peripheral level. Informational masking (IM), on the other hand, is said to occur "when the signal and masker are both audible but the listener is unable to disentangle the elements of the target signal from a similar-sounding distracter" (Brungart, 2001, p. 1101). More generally, the term IM has been applied to a wide variety of auditory masking processes which may have nothing more in common than the fact that they do not appear to involve EM.

Various other factors determine the extent to which EM and IM are effective. When a masker fluctuates in level, EM can be reduced by "glimpsing" acoustic information during the momentary reductions in masker energy, also known as "dip listening" (Miller and Licklider, 1950; Howard-Jones and Rosen, 1993a). Howard-Jones and Rosen (1993b) made a distinction between two different kinds of fluctuations in masker energy. Sometimes maskers fluctuate uniformly across their spectrum, as in the case of an amplitude-modulated broad-band noise, giving rise to *comodulated glimpses*. In other maskers, glimpses can be *uncomodulated*, meaning that they are restricted in frequency at any one time. Many natural masking signals, such as speech, have mixtures of both. For example, a comodulated glimpse will arise during the silent interval preceding the release of a voiceless plosive between vowels ("*the pack*"). Uncomodulated glimpses occur during the natural variations in spectrum across time that are characteristic of speech and essential for its intelligibility (Rosen and Iverson, 2007).

Consider how opportunities for glimpsing change as the number of talkers in the masker increases. For a masker consisting of a single talker, EM will only occur when energy in the masker coincides in spectrum and time with the target speech, and is sufficiently intense in that region to reduce the target's audibility. As the number of talkers in the background increases, the overlapping energy of those talkers will fill in the spectro-temporal dips in the masker and reduce the opportunities for glimpsing. Hence, we expect the EM effect of such a masker to increase monotonically with the number of talkers the masker contains, until the number of talkers becomes large enough so that adding more has no further effect. The most effective energetic masker, then, should be a broad-band noise with a spectrum shaped to that of the target speech because such a noise has minimal fluctuations in which to glimpse.

Note though that a recent study by Stone *et al.* (2012) complicates a simple interpretation of release from EM. They argued that much of what has been labeled EM is, in fact, more related to the masking of *modulations* rather than energy directly, and that glimpsing is only beneficial in maskers that are modulated. Although an important issue generally, insofar as all our maskers will be modulated in various ways at the outputs of auditory filters, it is not crucial whether the release of masking relates more to modulations than to energy. Furthermore, it is not yet clear how these ideas apply to complex periodic maskers. Therefore, for simplicity, we only refer to EM and release from it, whilst acknowledging the likelihood that modulation masking may also be important.

IM, as opposed to EM, is thought to vary according to the similarity between the target speech and the masker. Because it is supposed to depend upon quite different, and more central, processes, the amount of IM probably will *not* change with the number of talkers in the masker in the simple way noted above for EM. Instead, adding more background talkers may *improve* target speech intelligibility because the background becomes less similar in percept to the target. Similarly, increasing the number of talkers makes individual words in the babble less detectable, thus reducing lexical interference (Hoen *et al.*, 2007).

More recent theorizing suggests that there are probably at least two different aspects of IM (Shinn-Cunningham, 2008). One of these arises from a listener's inability to separate two (or more) distinct auditory objects, or perform the appropriate "auditory scene analysis" (Bregman, 1990). This kind of IM appears to be the kind most often referred to, and is perhaps exemplified when the masker is a single talker of the same sex and similar voice quality to the target. But maskers are also able to interfere with the perception of a target by pulling attention away from it, in essence by distraction. Shinn-Cunningham (2008), drawing on the visual perception literature, labeled these two distinct phenomena as "object formation" and "object selection."

One crucial contributor to the degree of IM related to auditory scene analysis, at least for maskers with small numbers of talkers, concerns aspects of periodicity and aperiodicity. Speech targets are typically quasi-periodic, so it might be thought that this periodicity is exploited by listeners to "enhance" the speech signal. Yet a number of studies appear to demonstrate that periodicity in the masker is of considerably greater advantage to a listener than periodicity in the target, leading to the notion that there is a cancellation (rather than enhancement) mechanism that depends upon harmonicity (e.g., de Cheveigne *et al.*, 1995). Vestergaard and Patterson (2009) disputed this notion on the basis of an experiment in which consonant-vowel and vowel-consonant syllables served as targets and maskers. Both targets and maskers could be synthesized with the ordinary excitation sources (preserving voicing) or with aperiodic excitation only (simulating whispered speech). They concluded that "listeners use voicing whenever it is present, either to detect the target speech or to reject the distracter" (Vestergaard and Patterson, 2009, p. 2863). In their view, it is the *difference* in sound quality arising from the presence or absence of periodicity that is important. Note that the claims about the usefulness of periodicity in both these studies apply to all maskers, whether modulated or not.

A more specific claim about the role of periodicity is that effective glimpsing requires access to differences in temporal fine structure between targets and maskers (Lorenzi *et al.*, 2006). Although this point has typically been made in the case of a genuinely aperiodic (noisy) masker, it appears likely that such a mechanism could operate even when the masker was periodic, because the target and masker could be distinguished on the basis of typically different fundamental frequencies (F0s). Note too that although

there is general agreement about the utility of periodicity generally for speech in noise, its relevance for glimpsing specifically has been disputed (Moore, 2011).

Here we attempt to clarify the role of the various factors mentioned in determining the intelligibility of speech in a competing babble of other talkers as the number of talkers in the babble changes. We used meaningful sentence targets because of the possibility that the trends observed previously in words, VCVs, and nonsense sentences might not hold for more ecologically relevant materials. The periodicity and aperiodicity of the targets and maskers were manipulated separately from other aspects related to IM and EM by using both targets and maskers that were noise-vocoded (Shannon *et al.*, 1995).

The use of noise-vocoded targets and signals can also offer insight into listening in noise by cochlear implant users who have little or no access to F0 information with current processing schemes. The temporal periodicity cues which can provide some F0 information in quiet (Green *et al.*, 2004; Souza and Rosen, 2009; Souza *et al.*, 2011; Arehart *et al.*, 2011) are insufficient to allow segregation of two signals (Qin and Oxenham, 2003; Stickney *et al.*, 2004; Arehart *et al.*, 2011).

Also, by including a steady-state speech-spectrum noise masker which was modulated by the wide-band envelope of the different babbles (as did Simpson and Cooke, 2005), we could determine the extent to which glimpsing was exploited, insofar as this signal varies only in glimpsing opportunities as the number of talkers it is based on changes. These conditions provide a kind of baseline from which to consider the degree to which glimpsing opportunities change in natural babble as the number of talkers varies.

In the first experiment, the number of talkers and type of masking noise were varied in a sentence recognition task using two fixed signal-to-noise ratios (SNRs) and natural speech targets. In the second experiment, we investigated the usefulness of periodicity cues by noise-vocoding the target, the masker, or both, and measured performance using an adaptive paradigm.

II. EXPERIMENT 1

A. Listeners

Listeners were 16 adults (4 males and 12 females) who spoke British English as their only or primary language and had no known hearing loss. Their ages ranged from 19 to 36 years, with a mean of 26 years. Approval for this study was obtained from the UCL Research Ethics Committee and informed consent was obtained from each listener.

B. Stimuli

The target stimuli were IEEE sentences (Rothauser *et al.*, 1969) spoken by an adult male British English talker. The sentences were grouped in 10-sentence lists, with similar phonetic content. Scores were based on five key words per sentence.

Four masking conditions were used: speech babble, noise-vocoded babble, speech-envelope modulated noise, and unmodulated noise. All maskers were created from recordings in the EUROM database of English speech (Chan *et al.*, 1995), consisting of different speakers reading 5- to 6-sentence passages. Sixteen male talkers were chosen on the basis of having a similar speaking rate, a standard British accent, and voice quality similar to that of the target talker. Passages were digitally edited to delete pauses of more than 100 ms. The result was a sound file approximately 21 s in duration for each talker, without any significant pauses. These were normalized to a common root-mean-square (RMS) and were the basis for all maskers.

Speech babble was created as follows. For the single-talker condition (referred to as "1-talker babble" for convenience), 1 of the 16 background talkers was randomly chosen. To create the 2-, 4-, 8-, and 16-talker conditions, the appropriate number of additional randomly selected talkers was digitally added to the talker(s) already present in the previously constructed condition. The spectrum of each babble was then equalized to the long-term average spectrum of the 16-talker speech babble.

To create noise-vocoded babble, each of the five previously constructed babbles (not individual voices) was processed using locally developed MATLAB software. Each of the babbles was digitally filtered into 12 bands, using sixth-order (three orders per side) Butterworth infinite impulse response filters. Filter spacing was based on equal basilar membrane distance (Greenwood, 1990) across a frequency range of 0.1–11 kHz. The output of each band was full-wave rectified and low-pass filtered at 30 Hz (fourth-order Butterworth) to extract the amplitude envelope. The cutoff was set this low to preclude the appearance of quasi-periodic fluctuations in the envelopes arising from quasi-periodic voiced speech (Rosen, 1992). The envelope was then multiplied by a wide-band noise carrier. The resulting signal (envelope $\times$ carrier) was filtered using the same bandpass filter as for the first filtering stage. The RMS level was adjusted at the output of the filter to match the original level in that band, before the signal was summed across bands.

To create the speech-envelope modulated noise, the envelope of each babble wave was extracted by full-wave rectification and low-pass filtering at 30 Hz. The envelope was multiplied by a broad-band noise which had the long-term average spectrum of the 16-talker babble. The unmodulated noise consisted of a broad-band noise shaped to the long term average spectrum of the 16-talker babble.

The speech targets were presented at two different SNRs (-2 and -6 dB) for each masker. These ratios were chosen on the basis of pilot studies in order to minimize floor and ceiling effects across conditions. Performance was thus measured in 32 conditions: 3 masker conditions (speech babble, noise-vocoded babble, modulated noise) $\times$ 5 numbers of talkers (1, 2, 4, 8, 16) $\times$ 2 SNRs (-2 dB, -6 dB), plus unmodulated noise at two SNRs.

C. Procedure

Each experimental session took place in a quiet room in which only the experimenter and the listener were present. Stimulus presentation and scoring were performed using

custom MATLAB software on a laptop computer. The stimuli were presented over Sennheiser (Wedemark, Germany) HD 25-1 headphones with the listener repeating back the sentence heard. The experimenter then scored which of the five key words were correctly perceived. No feedback was given, but a listener was occasionally reinforced or corrected as to whether they had attended to the correct talker in babble masker conditions.

The target stimuli were between 1.9 and 2.4 s in duration. A randomly selected segment of the available 21 s of a particular masker was added to the target stimulus such that the masker began 400 ms earlier and finished 200 ms later than the target. For each presentation, the noise level was fixed at 70 dB sound pressure level (SPL, no weighting) over a frequency range of 0.1–5.0 kHz (as measured on a B&K Artificial Ear type 4153, Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark) while the target level was adjusted to achieve the specified SNR. The order of the conditions followed a randomized Latin square. Two lists (20 sentences) were presented for each condition. Before data collection began, the listener was familiarized with the task by responding to a set of 10 practice sentences which included 3 sentences in quiet and 7 sentences in which the target stimuli were combined with a variety of masker conditions.

D. Statistical methods

Although some aspects of the results can be addressed through straightforward analyses of variance (ANOVAs), some important questions concern trends in performance as the number of the talkers in the babble varies. A preliminary inspection of the data (Fig. 1) revealed non-monotonic changes for the speech masker, and changes not well described by a simple straight line for the other two maskers. Therefore, we used a technique known as *segmented regression* (Ritz and Streibig, 2008), in which it is assumed that the data can be fit by two straight lines of arbitrary slopes with a breakpoint at which the lines meet. This requires five parameters (two slopes, two intercepts, and the breakpoint) but the constraint that the two lines must meet at the breakpoint means that only 4 parameters need to be estimated. The regression was done separately for each masker condition but with both SNRs in a single model, meaning that the saturated model had 8 parameters. Standard statistical methods using F-tests on nested models were used to minimize the number of free parameters necessary to describe the results. Lines whose slopes were not significantly different from zero at the 0.05 level were set to zero. A logarithmic scale for talker number was assumed for these fits, excluding the data for unmodulated speech-shaped noise. Comparing parameter estimates across masker types was done using 85.6% confidence intervals as recommended by Payton *et al.* (2003) in order to maintain the $p \leq 0.05$ level when standard errors of the two estimates do not differ by more than a factor of 2, which was typical in the resulting models.

E. Results

Figure 1 shows boxplots of the results for the two SNRs separately along with the fits from the three segmented regression models. All masker conditions were better fit by two lines than by one, but no condition required the saturated model of 8 parameters. The number of talkers at which the breakpoint occurred and the slopes of the lines for higher talker numbers were statistically indistinguishable across the two SNRs for each masker condition, so were required to be equal. In the case of the speech babbles, the breakpoint was fixed at the minimum performance level of 2 talkers because the fit was degenerate (i.e., any breakpoint in the range between 1 and 2 talkers led to equally good fits). Only for the modulated noise maskers was there no evidence of a change in performance after the breakpoint (as shown by the horizontal lines between about 5 and 16 talkers). Table I shows the parameters obtained. Note that the values for the

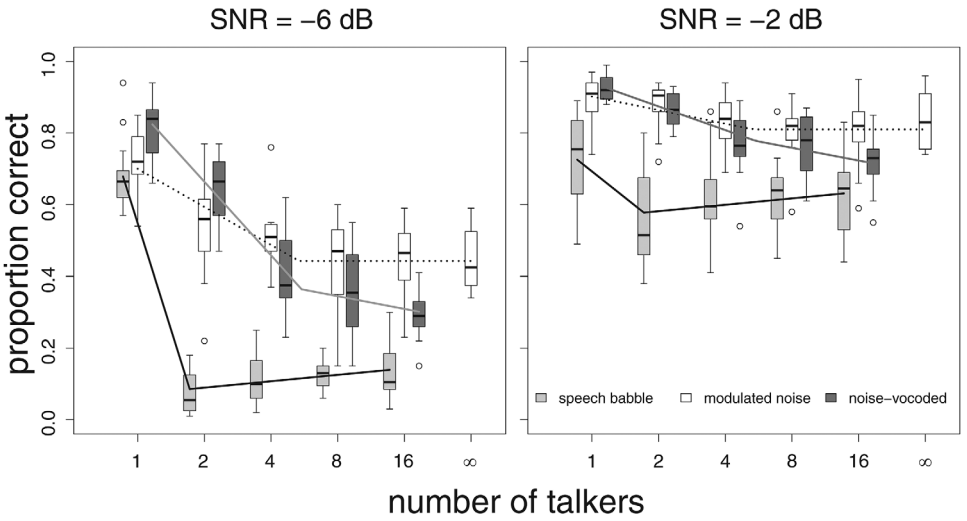

FIG. 1. Boxplots of the results obtained from Experiment 1 for the two SNRs separately in three different kinds of masker. Results for the speech-shaped unmodulated noise are plotted as occurring for an infinite number of talkers in the modulated noise. Also shown are the best fit lines from three segmented regressions, one for each noise type but including both SNRs. The prediction for the modulated noises has been extended to the results for the unmodulated noise even though that data was not used in the fits. Note that the staggering of the boxplots means that the *x*-axis values only strictly apply to the middle boxes, those concerning the results with modulated noise maskers. The other results have been shifted horizontally.

TABLE I. Selected parameter estimates for the segmented regression models used to provide the fits shown in Fig. 1. Slopes are given separately for the line below the breakpoint ("lower slope") and above ("upper slope"). Breakpoints have been transformed from their logarithmic values back to number of talkers for ease of interpretation. Also given are the proportions of variance accounted for (R^2) by each model (with 5 or 6 parameters), as well as the proportions accounted for by a simple ANOVA (10 parameters), using the R^2_1 recommended by Kvålseth (1985). In no case did the segmented regression model fit the data statistically worse than a simple ANOVA [$F(5,155) = 0.47$, $p = 0.80$; $F(5,155) = 1.68$, $p = 0.14$; $F(4,154) = 1.07$, $p = 0.38$ for the babble, modulated noise, and noise-vocoded maskers, respectively].

	Lower slope					
	SNR = −6 dB	SNR = −2 dB	Common upper slope	Common breakpoint	R^2 regression	R^2 ANOVA
Babble	−0.59	−0.15	0.02	2.0	0.88	0.88
Noise	−0.11	−0.04	0.00	5.4	0.78	0.80
Vocoded	−0.21	−0.07	−0.04	4.7	0.87	0.88

slopes of the lines directly indicate the change in proportion correct for a doubling of talker number. All three models accounted for the data statistically as well as a simple ANOVA model, as can also be seen by the proportions of variance accounted for by them in the two cases. The trends in the results were very similar for the two SNRs but were much clearer for SNR = −6 dB as there is a much greater range of performance. We therefore focus on that set of conditions.

Perhaps the most notable aspect of the results is that for a given number of talkers, other speech is always the most effective masker, as Miller (1947) surmised. One-way repeated-measures ANOVAs at each number of talkers shows masker type to be a significant factor in all five comparisons [$F(2,30) \geq 16.0$, $p < 0.001$ uncorrected]. Direct contrasts of performance with unprocessed babble against the other two maskers again show highly significant differences for all numbers of talkers [$F(1,15) \geq 20.0$, $p < 0.001$ uncorrected] except for modulated noise at 1 talker, where statistical significance would not survive a correction for multiple comparisons [$F(1,15) = 4.8$, $p = 0.045$ uncorrected].

Considering first performance for speech babble maskers, scores clearly decreased in going from 1 to 2 background talkers with a minimum in performance at 2 talkers. There was a slight but significant improvement in scores between 2 and 16 talkers as revealed in the segmented regression model.

For modulated noise, scores decreased almost monotonically as the number of background talkers increased, but with little change after 4 talkers. In the segmented regression model, this is reflected in the estimated breakpoint of 5.4 talkers, after which there is no statistical evidence for a change in performance for more talkers.

For the noise-vocoded babble, scores decreased more uniformly, and at a greater rate as the number of background talkers increased than for modulated noise. There is also statistical evidence that performance continued to decrease after the breakpoint of 4.7 talkers.

Considering trends across masker types, in all cases the greatest changes in performance occurred as the number of talkers in the masker increased from 1 to 2 or 4. For talker numbers greater than 4, performance changed much more slowly, if at all. Additionally, masker type interacted strongly with the number of talkers in determining performance in that there were large differences for different maskers in the rate at which performance decreased as the number of talkers increased from one. The decrement in performance for speech babble maskers in going from 1 to 2 talkers was considerably greater than for the other two maskers, and that for the noise-vocoded maskers was greater than what was obtained for modulated noise in going from 1 to 4 talkers. In fact, the confidence intervals for the lower slopes for the three different masker types did not overlap, indicating that they are statistically different (babble: −0.64 to −0.55; vocoded: −0.23 to −0.18; noise: −0.13 to −0.08).

Confidence intervals calculated for the breakpoints (except for speech babble, because of the degenerate fit) indicated that speech babble had a lower breakpoint than the other two maskers, but that the latter did not differ (speech babble: 2, vocoded: 3.8 to 5.8, noise: 4.1 to 7.1).

Finally, maskers appeared to differ in how their effectiveness changed beyond their breakpoints. Comparing the confidence intervals of the upper slopes showed that these slopes differed for speech babble and noise-vocoded maskers (being opposite in sign) whereas the value for modulated noise overlapped with both other maskers (babble: 0.01 to 0.03, vocoded: −0.06 to −0.01, noise: −0.03 to 0.03).

F. Discussion

Speech babble was the most effective masker type at every talker number. Performance for all three masker types was relatively close for 1-talker maskers, diverging sharply with increasing talker number. Performance dropped off most steeply for speech babble, and least steeply for modulated noise. Speech babble reached its breakpoint at 2 talkers, whereas the two other maskers had their breakpoints significantly higher, at about 5 talkers. After the breakpoint, masker effectiveness appeared to change differently for the three masker types. Adding more talkers to noise-vocoded maskers decreased performance further, albeit at a slower rate, whereas adding talkers to speech babble *increased* performance. Modulated noises showed no significant changes in performance after the breakpoint.

Clearly, an important determinant of the pattern of results for all maskers concerns differences in the opportunities to glimpse, with dips in signal energy expected to decrease, both in depth and frequency, monotonically with the number of talkers on which the masker was based. This may, in fact, be the only factor that is important for

modulated noise. In that case, the pattern of performance with number of talkers can be readily understood as the result of energy minima (which, being comodulated, extend across the whole spectrum) being "filled in" to the extent of becoming unusable when five or more talkers were present in the babble. Note too that predictions made on the basis of 1–16 talkers also accounted well for the performance obtained with unmodulated noise (marked as ∞ in Fig. 1).

To explore the role of glimpsing in the other conditions, let us make the assumption that unmodulated noise interfered with speech perception only through EM. Better performance with any other masker must therefore reflect a release from EM larger than any increase in IM. For −6 dB SNR, this constraint was met in five conditions [all at $t(15) > 3.55$, $p \leq 0.003$, meeting a Bonferroni-corrected value of $p = 0.0033$]: modulated noise with 1 or 2 talkers (allowing comodulated glimpsing), vocoded noise with 1 or 2 talkers (allowing comodulated and uncomodulated glimpsing), and speech babble with 1 talker (again, allowing comodulated and uncomodulated glimpsing). All these decreases in EM arise, at least in part, from opportunities to glimpse.

But note that there may still be less EM in conditions with worse performance than with unmodulated noise, if the reduced EM has been overwhelmed by increases in IM. Consider, in this light, the results for speech babble as a masker: improving performance as the number of talkers increases above 2 talkers suggests a release from IM greater than the effects lost to reduced glimpsing. More generally, it is not surprising that listeners performed more poorly with speech babble than with modulated noise, as was also reported by Simpson and Cooke (2005), at least for talker numbers greater than two. As mentioned above, there is likely to be lexical interference from the babble, which would be exceptionally strong for small numbers of talkers. There is also a thorny question, impossible to resolve in this study, about the relative utility of uncomodulated and comodulated glimpses. The envelope for the modulated noise is calculated on the basis of energy across the entire spectrum. Therefore, comodulated fluctuations in energy in unprocessed speech will be more-or-less preserved in modulated noise, but uncomodulated fluctuations will be smeared across frequency into comodulated fluctuations of shallower depth, or may even disappear. It is not known if any remaining shallower comodulated fluctuations in the modulated noise are of greater or lesser use than the spectrally restricted deeper ones in the babble.

Noise-vocoded babble has much in common with unprocessed babble. The former is also intelligible (at least for 1–2 talkers) so might have caused lexical interference, and it allows similar glimpsing opportunities. Nonetheless, there was less masking from the noise-vocoded babble. It seems likely that this difference arises, in part, from the difference in sound quality between the target speech and noise-vocoded babble. Speech contains much quasi-periodic energy (so has a strong pitch) whereas the noise-vocoded versions are strictly aperiodic. This difference might allow the listener to segregate target speech from noise-vocoded maskers much more readily than target speech from natural speech maskers. Note that this claim is contradictory to the idea that it is harmonic *cancellation* that is crucial when considering aperiodic and (mostly) periodic maskers (de Cheveigne *et al.*, 1995). If harmonic cancellation was primary, then noise-vocoded maskers should be more effective maskers than ordinary speech. If however, it is a *difference* in sound quality (related to periodicity) that is important (as claimed by Vestergaard and Patterson, 2009), we should get similar results when the *target* is vocoded and the babble masker is not. That situation is explored in Experiment 2.

III. EXPERIMENT 2

A. Listeners

Listeners were 20 adults (14 female), who spoke British English as their first language and had no known hearing loss. They ranged in age from 19 to 61 years, with a mean of 28 years. Local Institutional Review Board procedures were followed and informed consent was obtained for each listener.

B. Stimuli

The stimuli were the same as for Experiment 1, with the following exceptions. In addition to the unprocessed sentences, a second target condition was created by vocoding the target sentences using the 12-channel noise vocoder described above. The masker conditions included speech babble, noise-vocoded babble, and modulated noise, as in Experiment 1. Because the largest changes in performance for all masker conditions occurred between 1 and 2 talkers, only the 1-, 2-, and 16-talker masker conditions were included, plus the unmodulated noise. Thus, the final set of experimental conditions consisted of 10 maskers × 2 targets, for a total of 20 conditions.

C. Procedure

The testing session took place in a quiet room with stimuli presented over Sennheiser HD-25-1 headphones. As before, the listener repeated back as much as possible of the IEEE sentence heard for scoring by the experimenter. Pilot testing indicated that it would be difficult to avoid floor and/or ceiling effects using the same fixed SNR for both vocoded and unprocessed targets. Accordingly, an adaptive procedure, based on Plomp and Mimpen (1979), was used to measure the speech reception threshold (SRT). Masker intensity was fixed at 65 dB SPL (measured in a B&K type 4153 Artificial Ear). The SNR was set at −20 dB for the first sentence of each condition, which was presented repeatedly with SNR increased in steps of 6 dB until the listener correctly identified all five key words. Following this, each sentence was only presented once. The SNR was increased when 0–2 key words were correctly reported and decreased otherwise, thus tracking 50% correct. An initial step size of 4 dB in SNR was reduced to 2 dB in equal dB steps over the next two reversals, with the adaptive procedure continuing until 20 sentences (2 IEEE lists) had been presented. SRTs were calculated as the mean of all reversals once the final step size had been reached.

For familiarization before testing, ten sentences were presented using noise-vocoded speech as the target and speech-shaped noise as the masker using the adaptive procedure. A further ten sentences were then presented with target and masker type matching the first condition in the actual

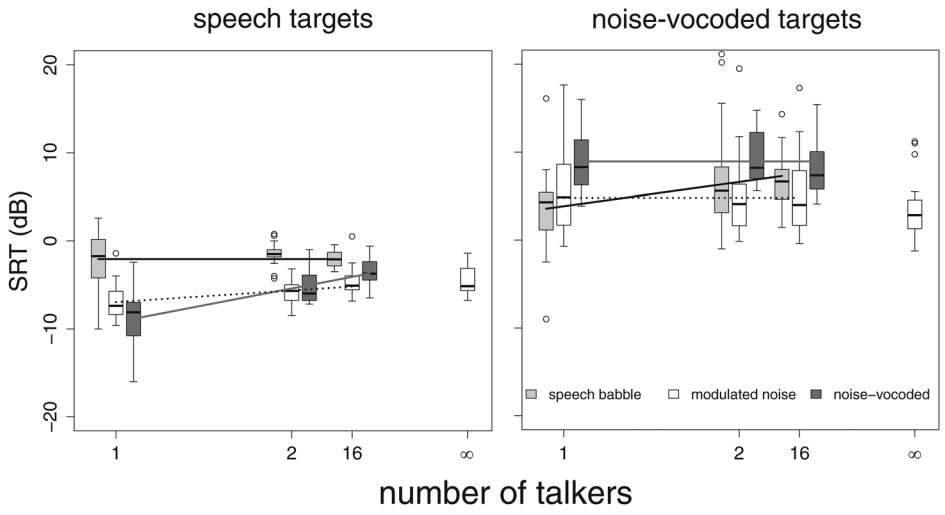

FIG. 2. SRTs as obtained in Experiment 2 shown separately for the two different target types in three different kinds of masker. Results for the speech-shaped unmodulated noise are plotted as occurring for an infinite number of talkers in the babble. Talker number has been transformed by a Box–Cox transformation with $\lambda = -1.95$, which led to the best fits in a linear regression for each target/masker combination (also shown). The slopes of the regression lines have been set to zero for the three target/masker combinations that led to slopes not statistically different from zero.

test. The 20 test conditions were then completed in an order determined by a randomized Latin square. No sentence list was repeated for any listener, and each pair of lists was heard exactly once in each condition.

D. Statistical methods

Again, although some aspects of the results can be addressed through straightforward ANOVAs, some important questions concern trends in performance as the number of the talkers in the babble varies. A preliminary inspection of the data (see Fig. 2) revealed monotonic changes in SRT as a function of talker number for all target/masker combinations. As there is no *a priori* way to predict in detail how SRTs will vary with talker number, an appropriate scaling of talker number was found using the Box–Cox family of transformations (Box and Cox, 1964). The Box–Cox transform has a single parameter which varies the mapping from expansive to compressive, including a logarithmic transform. The optimal fit of a saturated model assuming a linear relationship between SRT and transformed talker number (with a separate slope and intercept for each target/masker combination) applied to all the data (excluding only that for unmodulated noise) led to the value $\lambda = -1.95$, rather more compressive than a log.

E. Results and Discussion

Boxplots of the SRTs obtained are shown in Fig. 2, in separate panels for the two different target types. Also shown for each combination of masker and target type are best-fit regression lines for performance as a function of the number of talkers in the masker, where talker number has been transformed by the optimal Box–Cox transformation.

1. Differences between unprocessed and noise-vocoded targets

Although not unexpected (Qin and Oxenham, 2003; Stickney *et al.*, 2004), the most obvious finding is that the SRT was substantially better when the target was unprocessed speech than when it was noise vocoded. There was almost no overlap of the distribution of SRTs for any combination of masker type and number of talkers, except for speech maskers at 1 and 2 talkers. In all ten conditions, the mean SRTs were significantly higher for the noise-vocoded than for the speech targets [all 10 paired t-tests at $p < 0.001$, with $t(19) > 6.5$], with differences ranging from 6–18 dB (with a mean of 11 dB).

Although small differences in instantaneous levels might have been caused by spectral smearing across the width of each band in the vocoded target, such minimal differences in signal audibility (hence in EM) are unlikely to account for the overall 11 dB difference in SRT between the two targets. This is most clearly illustrated in the aggregate psychometric functions (PFs) for unprocessed speech and noise-vocoded targets in the unmodulated speech-shaped noise seen in Fig. 3. Other factors must be operating.

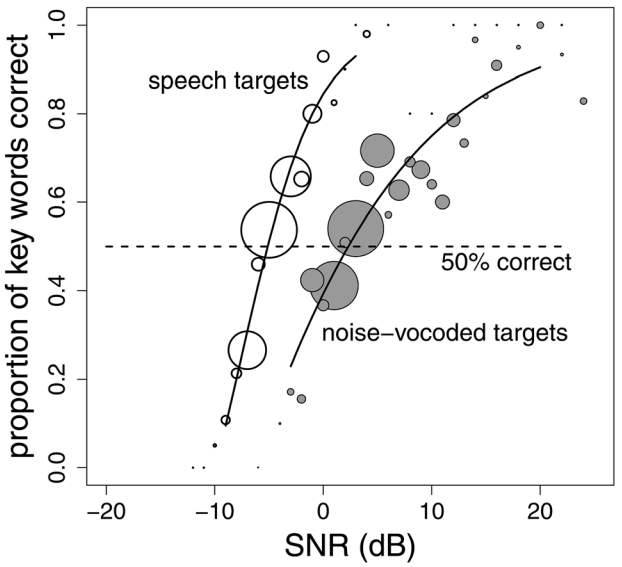

FIG. 3. Psychometric functions (PFs—the proportion of key words identified as a function of SNR) aggregated across all listeners repeating back both ordinary speech and noise-vocoded target sentences in the background of an unmodulated speech-spectrum noise. The size of the plotted circles indicates the number of trials at that particular SNR. Responses to the first sentence in each adaptive track (when the SNR was increasing from a low value) were eliminated, as were 2.5% of the trials from the high and low end of the distribution of SNRs. Note that performance for the noise-vocoded targets does not uniformly reach 100% even at the highest SNRs. Also shown are the results of two separate logistic regressions which fit sigmoid-shaped functions to each PF.

First, the noise-vocoded speech targets will be relatively unfamiliar. Although listeners were given practice with the task and stimuli, none had experience prior to this study in listening to noise-vocoded speech. Clear adaptation to noise-vocoded speech over tens of sentences has been demonstrated before (Davis *et al.*, 2005), but it is not known how much experience is necessary before performance reaches an asymptote. It is thus entirely possible that, even at the end of the experiment, listeners were not performing as well with vocoded targets as they might have with further training.

Second, and likely more importantly, noise-vocoded speech differs from natural speech in a number of ways likely to make it less intelligible. Intonation is eliminated, and the distinction between periodic and aperiodic excitation erased. Even a 12-channel vocoder will smear the spectral detail in the signal and the inherent fluctuations in the noise carrier will disrupt the envelope information within each auditory frequency channel. Although such changes are not sufficient to prevent perfect performance in quiet, at least for practiced listeners, impaired performance with noise-vocoded speech can be readily revealed in more difficult situations (Faulkner *et al.*, 2001; Friesen *et al.*, 2001).

This large difference in performance for the two types of target also has important implications for interpreting the results because it appears that the extent to which glimpses are useful decreases with increasing SNR (Bernstein and Grant, 2009; Bernstein and Brungart, 2011). Therefore, any differences in glimpsing across target type cannot be attributed only to the relationship between the acoustic properties of the target and masker.

2. Effects of the number of talkers in the masker

The effect of the number of talkers for each masker type was generally small. In order to characterize these effects more thoroughly, a linear mixed models analysis was applied for each target type separately, with masker type as a categorical predictor, and the Box–Cox transformed number of talkers as a continuous one. Also included in both models was the interaction term between masker type and number of talkers, as this assesses the extent to which the change in performance with number of talkers is different for different maskers. In fact, this interaction term was highly significant for speech targets [$F(2,174) = 12.7$, $p < 0.001$], but only marginally so for vocoded targets [$F(2,174) = 3.0$, $p = 0.053$]. This almost certainly is a result of the greater variability for the vocoded targets.

In order to clarify the nature of the interaction, separate linear mixed models were applied to each of the six target/masker combinations. For the noise-vocoded targets, increasing the number of talkers resulted in poorer SRTs only for the speech babble [$F(1,58) = 6.3$, $p < 0.02$ and $F(1,58) \leq 0.1$, $p > 0.75$ for the other two maskers]. For the unprocessed targets, exactly the complementary effects were obtained: only for the babble maskers did increasing the number of talkers *not* have a detrimental effect [$F(1,58) = 1.2$, $p > 0.25$, with $F(1,58) \geq 17.8$, $p < 0.001$ for the two aperiodic maskers].

Again we would expect glimpsing opportunities to decrease with the number of talkers, so any improvements in performance as talker number decreases may well reflect the exploitation of glimpses. For the noise-vocoded targets, the absence of any evidence for glimpsing with either aperiodic masker is consistent with the results of cochlear implant simulations, which also reveal little or no glimpsing (Nelson *et al.*, 2003; Cullington and Zeng, 2008). In those studies, the target speech and masker are mixed together before the noise-vocoding, but the resulting stimuli would be quite similar to those used here. Because glimpsing does not seem to be a factor for either aperiodic masker, presumably the difference in performance between them (about 4 dB more masking for the noise-vocoded masker) results from differences in IM. For small numbers of talkers (1 and 2), words are clearly audible in the noise-vocoded maskers. But even for 4–16 talkers, there is still an impression of people talking, however unintelligibly. The percept is quite different from the modulated noise maskers, even though they themselves are fluctuating more than the steady-state speech noise. Presumably this impression arises from the fact that the noise-vocoded maskers always have at least some variations in spectrum over time, derived from natural speech, which the modulated noise maskers do not.

Interestingly, performance in speech babble maskers for the noise-vocoded targets *did* improve as talker number decreased, which could result from more opportunities to glimpse. This might be seen as confirmation of the idea that effective glimpsing requires access to differences in temporal fine structure between targets and masker (Lorenzi *et al.*, 2006). It is also interesting that this purported glimpsing occurs at high SNRs, against the claims of Bernstein and Grant (2009) that glimpsing does not occur for SNRs > 0. On the other hand, the improvement with decreasing talker number could also reflect the fact that there are fewer F0 contours to track and cancel (de Cheveigne *et al.*, 1995; Hawley *et al.*, 2004).

For the unprocessed speech targets, improvements in performance with fewer competing talkers for the aperiodic maskers can perhaps be more clearly claimed to result from better opportunities to glimpse, because there are no complications arising from changes in the number of F0 contours. Also, the SNRs obtained here were low, consistent with Bernstein and Grant's claims that glimpsing is more effective at low SNRs.

3. Comparing trends in results across Experiments 1 and 2

Perhaps the most unexpected finding was that the SRT for a speech target in speech babble did not change with talker number. In Experiment 1 at SNR $= -6$ dB, performance plummeted as the number of talkers in the babble went from 1 to 2. More generally, the number of talkers in the masker appeared to have much smaller effects on SRT than it did on performance in Experiment 1.

These apparent discrepancies can be resolved through consideration of the PFs. Responses were aggregated across all 20 listeners, for all trials after the first sentence (which was presented multiple times), separately for each combination of target and masker. For each PF, logistic regression was used to estimate the best-fitting sigmoid function to the

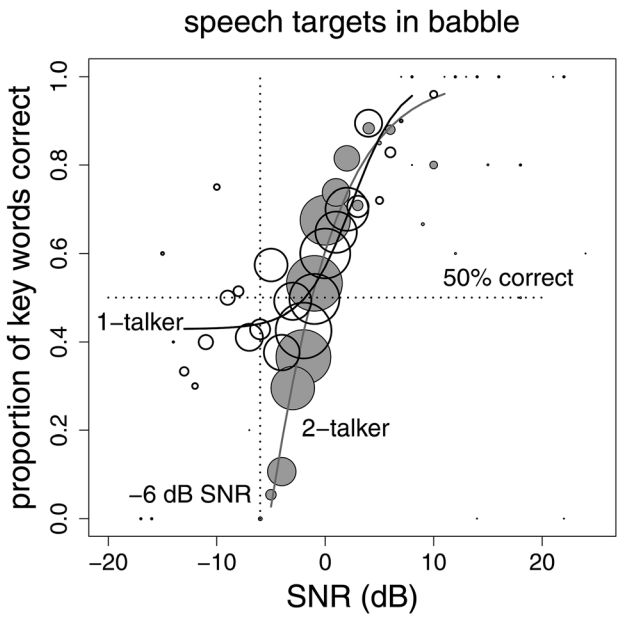

FIG. 4. PFs aggregated across all listeners repeating back unprocessed speech target sentences in the background of 1- and 2-talker babble. Responses to the first sentence in each adaptive track (when the SNR was increasing from a low value) were eliminated, as were 2.5% of the trials from the high and low end of the distribution of SNRs. The size of the plotted circles indicates the number of trials at that particular SNR. Also shown are the results of two separate logistic regressions which fit sigmoid-shaped function to each curve. Only the results for the 1-talker babble masker required the use of a lower plateau to the fitted sigmoid.

data. Only the central 95% of the distribution of SNRs was included in this analysis, trimming 2.5% off each end, because such fitting procedures tend to be very sensitive to small changes in the tails of the distributions. We also allowed for the possibility of a lower plateau of performance significantly above zero (Ritz and Streibig, 2005). In fact, only one target/masker combination required a non-zero lower plateau—a speech target in 1-talker babble, shown in Fig. 4 along with the PF for speech in 2-talker babble. Here, it can be clearly seen that the PFs for the two conditions are very similar for performance levels greater than 50% or so—hence, the adaptive procedure results in very similar SRTs. The two curves diverge strongly for SNRs lower than about −2 dB. For the 2-talker babble, performance decreases sharply with worsening SNRs, but, for 1-talker, plateaus at about 40% even for SNRs down to −10 dB. Such plateaus are frequently seen in conditions when a single talker is the masker (Brungart, 2001; MacPherson, 2013), and clearly account for the different trends for these conditions across Experiments 1 and 2.

In order to determine more fully the extent to which the results from Experiments 1 and 2 were similar, expected performance for the levels in Experiment 1 (SNRs = −2 and −6 dB) were calculated from the logistic regressions to the PFs in Experiment 2 from the ten relevant conditions. These are plotted in Fig. 5, for comparison with the results from Experiment 1 in Fig. 1. Apart from somewhat lower performances overall, the patterning of results across the two experiments is very similar.

Note too, that only one PF required this lower plateau, all the others having the expected sigmoid shape. Furthermore, the slopes of the PFs were similar within target type, although steeper for the speech targets than the noise-vocoded ones (as can also be seen in Fig. 3). These properties make it meaningful to compare SRTs across masker condition and talker number, as long as the comparisons are done within target type where SRTs were reasonably similar. As noted above, comparing across target types may not be meaningful because of the dependence of glimpsing on overall SNR (Bernstein and Grant, 2009; Bernstein and Brungart, 2011).

4. Differences between masker types

Apart from the effect of talker number, a crucial outcome concerns the relative effectiveness of the different types of masker. From Fig. 2, it is clear that the greatest masking effect always occurred for the masker type most similar to the target (which was true for every number of talkers in the masker). Target/masker similarity has, in the view of some (e.g., Brungart, 2001), been the defining feature of IM, so it would seem that aspects of IM are the most crucial in this outcome.

Because the interaction in the linear mixed models analyses can make the interpretation of main effects misleading, a repeated-measures ANOVA was used to compare masker

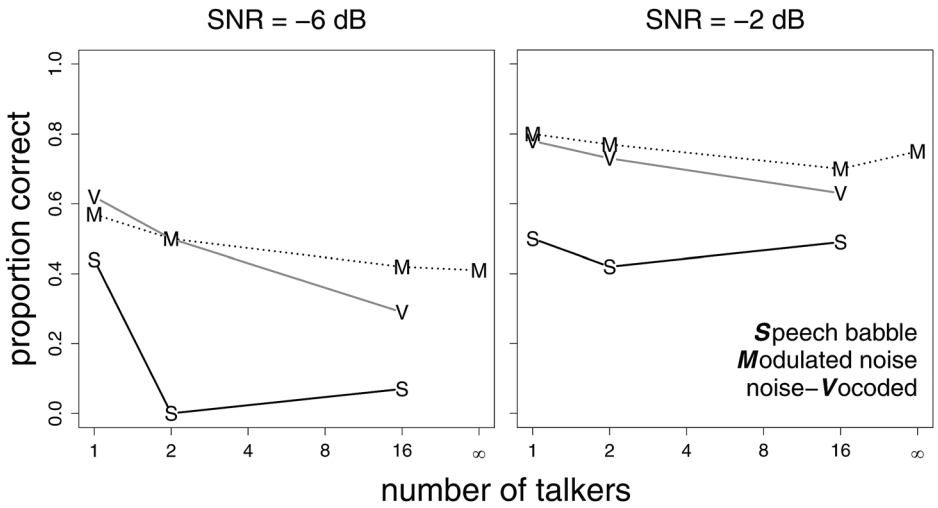

FIG. 5. Levels of performance under the fixed SNR conditions in Experiment 1 as predicted from logistic regressions of the aggregated results of Experiment 2. Only for the 1-talker babble masker was a lower plateau to a fitted sigmoid used. Note the general similarity in the pattern of results to those portrayed in Fig. 1.

types for each number of talkers and target type. In addition, a simple contrast was used to compare performance on the masker type identical to the target to the other two maskers.

For the speech targets (Fig. 2, left panel), there were highly significant differences in performance across the three maskers for all talker numbers [$F(2,38) > 18.8$, $p < 0.001$]. In each case, performance with the speech masker was significantly worse than for the other two maskers [$F(1,19) > 12.9$, $p \leq 0.002$]. In order to quantify the relative masking effectiveness of the two aperiodic maskers, a linear mixed models analysis was applied only to those two conditions. A significant interaction term [$F(1,116) = 12.6$, $p \leq 0.001$] showed that performance changed more with talker number for the noise-vocoded maskers than for the modulated noises, just as it did in Experiment 1. This interaction is perhaps not surprising. One way to think about the effects of adding talkers to a babble is as a kind of smoothing of fluctuations in energy. Noise-vocoded maskers will have both their comodulated and uncomodulated fluctuations smoothed as more talkers are added, whereas modulated-noise maskers will only have comodulated fluctuations smoothed. In other words, there is less smoothing to do to the latter as the original construction of the modulated noises already includes strong smoothing across frequency.

Consider now the noise-vocoded targets (Fig. 2, right panel). There were highly significant differences in performance for the three maskers for 1 and 2 talkers [$F(2,38) > 10.5$, $p < 0.001$], and in those cases, performance with the noise-vocoded masker was significantly worse than for the other two maskers [$F(1,19) > 10.8$, $p < 0.005$]. For 16 talkers, the effects were less marked. The three maskers did differ in their effects [$F(2,38) = 4.9$, $p = 0.015$] but differences in performance for the speech and noise-vocoded maskers did not reach statistical significance [$F(1,19) = 3.5$, $p = 0.08$]. Noise-vocoded maskers led to statistically worse performance than modulated noises [$F(1,19) = 7.6$, $p = 0.013$].

Comparing only the modulated noise and babble maskers, the boxplots suggest that performance changed more with talker number for the babble maskers than for the modulated noises. This interaction almost reached statistical significance [$F(1,116) = 3.6$, $p = 0.06$]. The tendency was for performance with the babble masker to be similar or slightly better than for the modulated noises for 1 talker, and then for babble to become a more effective masker than the noises at 16 talkers. Here again is evidence against the idea that cancellation of a periodic masker is a crucial factor. If cancellation were important, we would expect SRTs for the noise-vocoded targets to be somewhat lower with the babble maskers than with modulated noise, which they clearly were not. But it is also surprising that babble seemed to be a more effective masker at 16 talkers because of the clear quality difference between the obviously "buzzy" babble and the noise-vocoded targets.

IV. GENERAL DISCUSSION AND CONCLUSIONS

A. Energetic masking and its release through glimpsing/dip listening

To understand the differences in effectiveness of the different maskers, we can think of every masking situation as a net effect of EM plus IM (Agus *et al.*, 2009). In this study, the unmodulated noise represented the closest approximation to "pure" EM as it contained no informational components. It also offered minimal opportunities for release from EM through glimpsing. Although the unmodulated noise level does not change over time in a specific frequency region, the *target* levels themselves are varying. Therefore, even at equivalent signal and masker intensities, speech peaks will sometimes exceed masker levels and as such there will be points in time where EM is reduced.

A release from EM can occur when the masker has a dip in energy sufficiently large in the spectro-temporal domain to survive smoothing through the auditory filters. All of the maskers used here (except for steady-state noise) allowed comodulated glimpsing, resulting from energy dips across the whole frequency range, although these would decrease as the number of talkers in the masker increased. Modulated noise, with its unvarying spectral shape, had only comodulated dips. This comodulation across all frequencies may also serve as a segregation cue (Qin and Oxenham, 2003), further reducing the masking effect of modulated noise. Speech babble and noise-vocoded babble also had spectrally restricted dips, which thus allowed uncomodulated glimpsing.

Can energetic considerations fully explain the difference between the masking effectiveness of speech babble and noise-vocoded babble? The spectrally restricted dips present in the original speech babble will be reflected in the noise-vocoded babble, but the depth of these will be limited by the smearing across frequency exacted by the relatively wide-band filters of the vocoder in comparison to auditory filters. Accordingly, the noise-vocoded babble would have less spectral detail—and shallower restricted-frequency fluctuations—than the babble. However, previous studies show a glimpsing advantage only for relatively large spectral gaps (Howard-Jones and Rosen, 1993b; Peters *et al.*, 1998), so it is unclear whether the loss by smearing of smaller spectral gaps will be meaningful. With regard to comodulated glimpses, the noise-vocoded babble and babble should be quite similar. In terms of EM, then, we would expect the two masker types to have relatively similar fluctuations in energy, hence to exert similar amounts of EM.

Both comodulated and uncomodulated dips would be expected to become less prevalent as the number of talkers in the babble increases. For example, with a speech target, performance in babble decreased sharply between 1 and 2 talkers, as was reported by Freyman *et al.* (2001). Presumably, adding a second talker fills many of the large energy dips in the speech of a single talker, increasing EM, although having multiple F0 contours may also be a factor for ordinary speech. In our data, performance for all maskers changed little above 4 talkers, where energy dips were probably reduced to a point where they offered no glimpsing advantage compared to unmodulated noise.

Other studies have demonstrated a plateau or breakpoint in performance at more than 4 talkers but these differences almost certainly were influenced by the nature of the target speech material. For example, Simpson and Cooke (2005) used VCVs, some of which (e.g., /ɑsɑ/) would be identifiable on the basis of a much shorter stretch of waveform in

comparison to the sentences used here. In that case, we would expect glimpses of quite short durations to be useful, and Simpson and Cook found performance reached a plateau when the modulated noise condition comprised 64 talkers. Using babble maskers, there was a breakpoint in performance with worst scores at 8 talkers, larger than the 6 talker breakpoint found by Miller (1947) with words, and the 2-talker one we found for sentences. Although this particular question was not the focus of their study, Freyman *et al.* (2004) also provide relevant data. They tested nonsense sentences in the background of babble maskers containing 2, 3, 4, 6, and 10 talkers (all female adults). Considering only the condition in which the target and masker came from a single loudspeaker in front of the listener, performance was found to be worse for 2 talkers, and then generally increased as more talkers were added.

In short, variation across studies in the number of talkers at which a breakpoint in performance occurs almost certainly depends heavily upon the nature of the target speech material. While testing with VCVs makes an interesting demonstration, everyday communication is more akin to a sentence recognition test.

B. Informational masking and masker similarity

In Experiment 1 (ordinary speech targets), intelligibility was lower for the babble masker than in either noise-vocoded babble or modulated noise, regardless of the number of background talkers. This occurred for 1-talker maskers (although differences were small), and was most noticeable with 2-talker maskers, with performance much poorer than with even an unmodulated noise. Given that dips in energy should be common when there are only a small number of talkers, these maskers should offer the greatest release from EM. Therefore, unprocessed babble must have an IM component strong enough to offset any advantage of the release from EM, an effect which must be strongest for the 2-talker babble. With larger numbers of talkers, EM will continue to increase and any IM should decrease as individual talkers "blend" to an overall percept with few distinctive features. However all the different factors are interacting, it is relatively easy to focus on 1 talker when only one other is in the background. But it appears to be much more difficult to attend to 1 talker when there are *2* competing talkers, at least in the situation studied here, in which all of the talkers, target and masker, were male and had similar voice quality.

As mentioned above, babble and noise-vocoded babble should have similar degrees of EM. Any IM attributable to semantic content should also be similar (although the unprocessed babble is likely to be more intelligible—and hence exert more IM—than vocoded babble, especially for more than 1 talker). Perhaps more importantly, if the cancellation of a periodic masker is important (de Cheveigne *et al.*, 1995), we would expect better performance for speech targets in babble. Yet performance was always better, sometimes dramatically so, for speech targets in noise-vocoded babble than in unprocessed babble. It seems likely that this difference can be attributed to a release from IM due to the difference in quality between the noise-vocoded babble and the target speech. The target is, of course, typically voiced, hence possesses a clear pitch, which the vocoded masker does not. Although previous work has established that differences in fundamental frequency contribute to release from masking (e.g., Brokx and Nooteboom, 1982), there has been curiously limited discussion of the importance of differences in quality in distinguishing a pitched signal in a noisy background, or vice versa. A notable exception is provided by Vestergaard and Patterson (2009), who showed that differences in periodicity (voiced vs "whispered" sounds) could be exploited by listeners to minimize masking, without regard to which was target and which masker. These results may also be more relevant to our experiment than those of de Cheveigne *et al.* (1995), insofar as de Cheveigne *et al.* contrasted harmonic complexes with inharmonic ones consisting of discrete spectral components, rather then genuine "noise."

Some other data support the idea that a quality difference may assist in segregating sounds. Freyman *et al.* (2001) found improved performance when they time-aligned a noise-vocoded version of a 2-talker babble with the babble itself, compared to the 2-talker babble alone. The combined signal would have more or less the energy fluctuations of the babble itself, but sound noisy instead of speech-like. In work by Arbogast *et al.* (2002), threshold SNR was about 20 dB worse when both speech and masker were sine-vocoded than when the masker was a noise, even though with the sine-vocoded masker there should have been no EM because the carrier bands of the signal and masker were alternated to prevent frequency overlap. Because such effects cannot be explained on the basis of greater EM, they point to some aspect of IM.

C. Untangling EM and IM effects in speech babble

Although varying the number of talkers in a speech babble has the advantage of ecological plausibility, too many aspects of this masker change at the same time to allow simple explanations of the masking effects. As a first step, we propose three main interacting effects that may operate in determining the relative contributions of EM and IM as the number of talkers in the masker increases: (i) Energetic masking will become more effective through the loss of opportunities for glimpsing; (ii) Informational masking arising from lexical interference, or the competition for neural resources that appears to go on even for unattended speech sounds (Scott *et al.*, 2004; Scott *et al.*, 2009), will decrease because the masker is less intelligible as talker number increases; (iii) Informational masking arising from the failure of auditory scene analysis based on tracking F0 contours will increase. However, with a sufficient number of talkers in the babble, the babble must become identical to a speech-shaped noise (although whether this ever occurs in real-life situations is unclear).

The use of aperiodic maskers and targets introduces further questions about the extent to which categorical differences in quality, arising from differences in periodicity, can be used by a listener. Of course, the noise-vocoded speech and babble-modulated maskers are meant to aid interpretation of

the other results by serving as simpler signals with at least some of the properties of the speech babble. It seems, however, that the use of maskers which are unintelligible but preserve the periodicity of the speech signal are likely to be informative (Deroche and Culling, 2011; Chen *et al.*, 2012). In this way, it should be possible to disentangle the effects of the number of F0 contours in the masker separately from the changes in glimpsing opportunities as the number of talkers increase.

ACKNOWLEDGMENTS

This work was supported by the Medical Research Council (Grant Number G1001255), the National Institutes of Health (Grant No. R01 DC60014), and the Bloedel Hearing Research Center. Many thanks to Michael Akeroyd, Josh Bernstein, and Tim Green for numerous suggestions on how to improve the original manuscript. Rich Freyman kindly provided details of his results from two published papers, and Christian Ritz generously provided statistical advice.

Agus, T. R., Akeroyd, M. A., Gatehouse, S., and Warden, D. (**2009**). "Informational masking in young and elderly listeners for speech masked by simultaneous speech and noise," J. Acoust. Soc. Am. **126**, 1926–1940.

Arbogast, T. L., Mason, C. R., and Kidd, G. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

Arehart, K. H., Souza, P. E., Muralimanohar, R. K., and Miller, C. W. (**2011**). "Effects of age on concurrent vowel perception in acoustic and simulated electroacoustic hearing," J. Speech Lang. Hear. Res. **54**, 190–210.

Bernstein, J. G. W., and Brungart, D. S. (**2011**). "Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio," J. Acoust. Soc. Am. **130**, 473–488.

Bernstein, J. G. W., and Grant, K. W. (**2009**). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 3358–3372.

Box, G. E. P., and Cox, D. R. (**1964**). "An analysis of transformations," J. R. Stat. Soc. Ser. B (Stat. Methodol.) **26**, 211–252.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (The MIT Press, Cambridge, MA), pp. 1–792.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Chan, D., Fourcin, A., Gibbon, D., Granstrom, B., Huckvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld, C., and Zeiliger, J. (**1995**). "EUROM—A spoken language resource for the EU," Eurospeech'95, in *Proceedings of the 4th European Conference on Speech Communication and Speech Technology*, Vol. 1, pp. 867–870.

Chen, J., Li, H. H., Li, L., Wu, X. H., and Moore, B. C. J. (**2012**). "Informational masking of speech produced by speech-like sounds without linguistic content," J. Acoust. Soc. Am. **131**, 2914–2926.

Cullington, H. E., and Zeng, F. G. (**2008**). "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects," J. Acoust. Soc. Am. **123**, 450–461.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. Gen. **134**, 222–241.

de Cheveigne, A., McAdams, S., Laroche, J., and Rosenberg, M. (**1995**). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. **97**, 3736–3748.

Deroche, M. L. D., and Culling, J. F. (**2011**). "Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation," J. Acoust. Soc. Am. **130**, 2855–2865.

Faulkner, A., Rosen, S., and Wilkinson, L. (**2001**). "Effects of the number of channels and speech-to-noise ratio on rate of connected discourse tracking through a simulated cochlear implant speech processor," Ear Hear. **22**, 431–438.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.

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

Green, T., Faulkner, A., and Rosen, S. (**2004**). "Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants," J. Acoust. Soc. Am. **116**, 2298–2310.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**, 833–843.

Hoen, M., Meunier, F., Grataloup, C. L., Pellegrino, F., Grimault, N., Perrin, F., Perrot, X., and Collet, L. (**2007**). "Phonetic and lexical interferences in informational masking during speech-in-speech comprehension," Speech Commun. **49**, 905–916.

Howard-Jones, P. A., and Rosen, S. (**1993a**). "The perception of speech in fluctuating noise," Acustica **78**, 258–272.

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

Kvålseth, T. O. (**1985**). "Cautionary note about R^2," Am. Stat. **39**, 279–285.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.

MacPherson, A. (**2013**). "The factors affecting the psychometric function for speech intelligibility," Ph.D. thesis, University of Strathclyde, Glasgow, Scotland, pp. 1–296.

Miller, G. A. (**1947**). "The masking of speech," Psych. Bull. **44**, 105–129.

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

Moore, B. C. J. (**2011**). "The importance of temporal fine structure for the intelligibility of speech in complex backgrounds," in *Speech Perception and Auditory Disorders*, edited by T. Dau, J. Dalsgaard, M. Jepsen, and T. Poulsen (Centertryk A/S, Denmark), pp. 21–32.

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

Payton, M. E., Greenstone, M. H., and Schenker, N. (**2003**). "Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance?," J. Insect Sci. **3**, 1–6.

Peters, R. W., Moore, B. C. J., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577–587.

Plomp, R., and Mimpen, A. M. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Audiology **18**, 43–52.

Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.

Ritz, C., and Streibig, J. C. (**2005**). "Bioassay analysis using R," J. Stat. Software **12**, 1–22.

Ritz, C., and Streibig, J. C. (**2008**). *Nonlinear Regression with R* (Springer, New York), pp. 1–148.

Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory, and linguistic aspects," Philos. Trans. R. Soc. London, Ser. B **336**, 367–373.

Rosen, S., and Iverson, P. (**2007**). "Constructing adequate non-speech analogues: What is special about speech anyway?," Dev. Sci. **10**, 165–168.

Rothauser, E. H., Chapman, N. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

Scott, S. K., Rosen, S., Beaman, C. P., Davis, J. P., and Wise, R. J. S. (**2009**). "The neural processing of masked speech: Evidence for different mechanisms in the left and right temporal lobes," J. Acoust. Soc. Am. **125**, 1737–1743.

Scott, S. K., Rosen, S., Wickham, L., and Wise, R. J. S. (**2004**). "A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception," J. Acoust. Soc. Am. **115**, 813–821.

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

Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cogn. Sci. **12**, 182–186.

Simpson, S. A., and Cooke, M. (**2005**). "Consonant identification in *N*-talker babble is a nonmonotonic function of *N*," J. Acoust. Soc. Am. **118**, 2775–2778.

Souza, P., Arehart, K., Miller, C. W., and Muralimanohar, R. K. (**2011**). "Effects of age on F0 discrimination and intonation perception in simulated electric and electroacoustic hearing," Ear Hear. **32**, 75–83.

Souza, P., and Rosen, S. (**2009**). "Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech," J. Acoust. Soc. Am. **126**, 792–805.

Stickney, G. S., Zeng, F. G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.

Stone, M. A., Fullgrabe, C., and Moore, B. C. J. (**2012**). "Notionally steady background noise acts primarily as a modulation masker of speech," J. Acoust. Soc. Am. **132**, 317–326.

Vestergaard, M. D., and Patterson, R. D. (**2009**). "Effects of voicing in the recognition of concurrent syllables," J. Acoust. Soc. Am. **126**, 2860–2863.