Consonant identification in *N*-talker babble is a nonmonotonic function of *N* (L)

Sarah A. Simpson[a] and Martin Cooke[b]

*Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street,
Sheffield, S1 4DP, England*

(Received 18 March 2005; revised 12 August 2005; accepted 18 August 2005)

Consonant identification rates were measured for vowel-consonant-vowel tokens gated with N-talker babble noise and babble-modulated noise for an extensive range of N, at a fixed signal-to-noise ratio. In the natural babble condition, intelligibility was a nonmonotonic function of N, with a broad performance minimum from $N=6$ to $N=128$. Identification rates in babble-modulated noise fell gradually with N. The contributions of factors such as energetic masking, linguistic confusion, attentional load, peripheral adaptation, and stationarity to the perception of consonants in N-talker babble are discussed. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.2062650]

PACS number(s): 43.71.Es, 43.66.Dc, 43.66.Lj, 43.50.Fe, 43.72.Dv [RLD] Pages: 2775–2778

I. INTRODUCTION

Speech communication frequently takes place in environments in which other talkers are active. For this reason, babble (i.e., the summed waveform of several simultaneous talkers) is often used as a masker in studies of everyday speech perception in noise. However, the masking effect of babble is heavily dependent on the number (N) of simultaneous talkers in the mixture. Recent studies employing babble noise, such as Snell *et al.* (2002), Markham and Hazan (2004) and Cutler *et al.* (2004) used $N=\{4,6,20\}$, respectively. The widely used speech intelligibility test of Kalikow *et al.* (1977) contains babble with $N=12$.

Single-talker maskers ($N=1$) and speech-shaped noise ($N=\infty$) are the extremes of the babble continuum. Speech reception threshold (SRT) gains of around 6–8 dB for the single-talker masker over speech-shaped noise have been reported (Duquesnoy, 1983; Festen and Plomp, 1990). The release from masking produced by a single talker relative to speech-shaped noise is usually explained by the assumption that listeners take advantage of temporal fluctuations in masker energy to listen in intervals of favorable local signal-to-noise ratio (SNR) (Assmann and Summerfield, 2004). If this were the only factor underlying speech perception in babble noise, one would expect that as N increases, intelligibility would decline monotonically to the level observed in speech-shaped noise.

Miller's classic study of masking (Miller, 1947) was the first to investigate intermediate values of N. Miller measured the intelligibility of words in N-babble for $N=\{1,2,4,6,8\}$. He found that the difference in masking effect for a single talker over two talkers was equivalent to an SRT difference of about 8 dB. Babble with $N=\{4,6,8\}$ produced an additional 3–4 dB of masking over the two-talker condition. Miller's results on their own indicate a monotonic decrease in intelligibility as N increases. However, taken together with those of Duquesnoy (1983) and Festen and Plomp (1990), they suggest that babble for $N=\{4,8\}$ is a more effective masker than speech-shaped noise. Other studies support this hypothesis. Danhauer and Leppler (1979) observed that consonants in a background of babble with $N=\{4,9\}$ talkers were recognized less well than in white noise at SNRs below 5 dB. Miller's own data suggest than the $N=8$ condition provides marginally less masking that $N=\{4,6\}$ for SNRs of 3 dB and below. Finally, in a pilot study, the authors found significantly greater masking of consonants in an $N=8$ babble condition than in speech-shaped noise at SNRs of 0, −6, and −12 dB. Bronkhorst (2000) summarizes data on speech intelligibility in multitalker backgrounds for $N<9$.

The purpose of the current study was to discover the shape of the intelligibility function for an extensive range of N values. The study was motivated by: (i) The possibility of a nonmonotonic change in intelligibility as N increases, (ii) the difficulty in comparing results from previous studies, as noted by Bronkhorst (2000), (iii) the relatively narrow range of N tested to date, and (iv) the observation that babble constructed using larger values of N is used routinely in speech

[a] Electronic mail: s.simpson@dcs.shef.ac.uk
[b] Electronic mail: m.cooke@dcs.shef.ac.uk

perception testing. In fact, a study by Carhart *et al.* (1975) did employ babble with a wide range of *N* values. They measured the intelligibility of spondees in *N*-talker babble for $N=\{1,2,3,16,32,64,128,\infty\}$. Their results appeared to confirm the suggestion of a nonmonotonic intelligibility function, but since their findings appeared only as an abstract, it is difficult to appreciate the precise pattern and significance of the results.

In the current study, listeners identified consonants presented in vowel-consonant-vowel (VCV) contexts in *N*-talker babble for $N=\{1,2,3,4,6,8,16,32,64,32,64,128,512,\infty\}$. In addition, intelligibility in babble-modulated speech-shaped noise was measured for the same values of *N*.

II. EXPERIMENT: CONSONANT IDENTIFICATION IN *N*-TALKER BABBLE

A. Stimuli

Speech stimuli were chosen from the VCV corpus collected by Shannon *et al.* (1999). Sixteen consonants (b, d, g, p, t, k, m, n, l, r, f, v, s, z, ʃ,tʃ) in the context of the vowel /ɑ/ were used. Two examples of each consonant from five male talkers were chosen, leading to a test set of 160 items. An additional 32 VCVs were used as practice items.

Speech stimuli were presented in 23 masking conditions consisting of *N*-talker babble (11 conditions), speech-shaped noise modulated by *N*-talker babble (11 conditions), and unmodulated speech-shaped noise. Babble stimuli were constructed from subsets of 1056 utterances spoken by 132 male talkers from dialect regions 1–3 of the TIMIT corpus (Garofolo *et al.*, 1992). TIMIT "shibboleth" sentences spoken by all talkers were not used. All utterances were normalized to have the same RMS energy prior to forming babble noise to ensure that they all contributed equally to the masker. Speech-shaped noise was created by processing white noise with a filter whose magnitude response was equal to the long-term magnitude spectrum of the entire set of sentences. Speech-shaped noise was multiplied by the envelope of *N*-babble waveforms to create *N*-babble-modulated speech-shaped noise. Following Brungart *et al.* (2001), the envelope was computed by convolving the absolute value of the base signal with a 7.2 ms rectangular window.

Masked VCVs were formed by adding a randomly selected fragment of masking noise to each consonant at a constant target-to-masker ratio of −6 dB. The masker and VCV were gated, i.e., started and stopped at the same time. Stimuli were presented at approximately 68 dB SPL.

B. Listeners

Twelve listeners (10 M and 2 F) participated in the experiment. All received a hearing test and were found to have normal hearing (better than 20 dB hearing level in the range 250–8000 Hz). All listeners passed a pretest which required them to recognize VCV tokens in clean conditions at an identification rate of at least 98%.

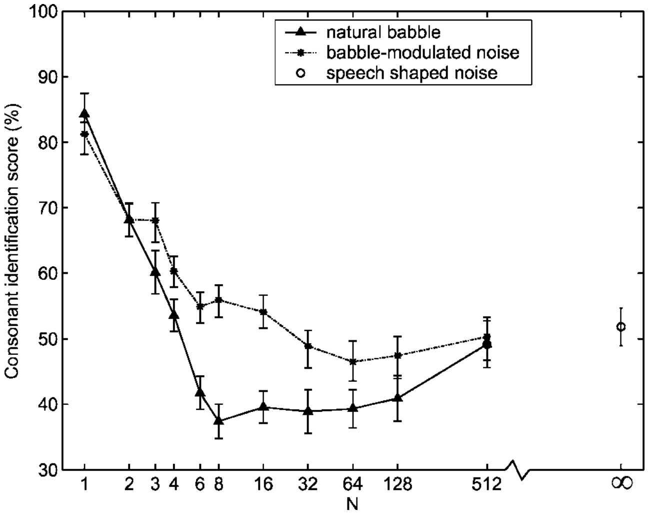

FIG. 1. Consonant identification rates in *N*-talker babble as a function of the number of talkers, for natural babble (solid line), babble-modulated noise (dashed line), and speech-shaped noise (circle). The error bars represent 95% confidence intervals in each condition.

C. Procedure

Listening sessions took place in an IAC single-walled acoustically isolated booth. Stimuli were resampled to 25 kHz and presented via a Tucker-Davis Technologies System 3 RP2.1. Stimulus presentation and results collection were controlled by a computer situated outside the booth. Signals were presented diotically over Sennheiser HD250 headphones.

Each participant completed the 23 conditions over 4–6 sessions. Every condition consisted of 192 tokens, and required about 6–7 minutes to complete. The initial 32 practice tokens were not scored, although participants were not aware of this. Condition orders were balanced across listeners, and token order within each condition was randomized.

III. RESULTS

Figure 1 summarizes consonant identification rates in all masking conditions. The data have been averaged across the 12 listeners and the error bars in the figure represent the 95% confidence interval at each data point. In natural babble, performance falls rapidly to a minimum at $N=8$. Little improvement is observed between $N=8$ and $N=128$ before a recovery to the level of speech-shaped noise by $N=512$. In contrast, babble-modulated noise is a less effective masker at all values of $N>2$ and shows a more gradual decrease in performance with increasing *N*. The difference between natural babble and babble-modulated noise also varied with *N*, reaching a maximum at $N=8$ (Fig. 2). The error bars in figure 2 represent the Bonferroni-adjusted 95% confidence intervals in each masking condition.

A repeated-measures ANOVA with factors of *N* and masker type showed a significant ($p<0.01$) effect for both factors and their interaction (effect size $\eta^2=0.924$ for masker type, 0.964 for *N* and 0.682 for their interaction). Results were partitioned by masker type and post-hoc tests (with Bonferroni adjustment for multiple comparisons) computed to investigate the effect of *N*.

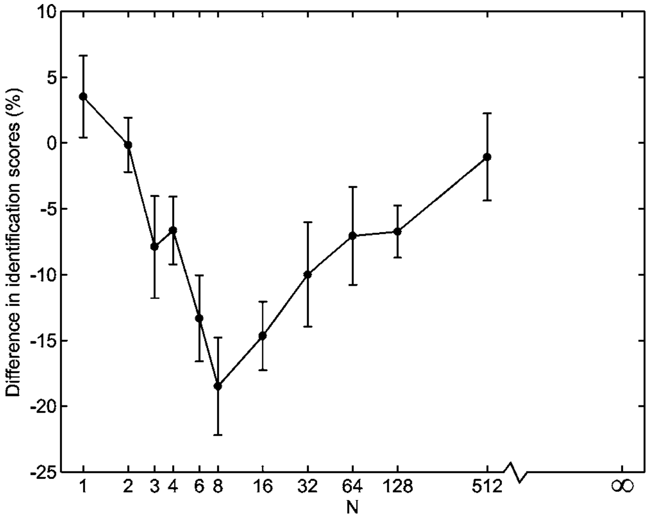

FIG. 2. Difference in consonant identification rate between natural babble and babble-modulated noise. The error bars represent Bonferroni-adjusted 95% confidence intervals in each condition.

For natural babble, conditions in the set $N=\{1,2,3,4\}$ differed ($p<0.01$) from each other and from all other values of N, apart from the pairs (3, 4), (4, 512), and $(4,\infty)$ which were not significantly different ($p>0.05$). All pairs of conditions in the subset $N=\{6,8,16,32,64,128\}$ were statistically equivalent. The $N=512$ condition differed from all others apart from $N=\{4,6,\infty\}$.

For babble-modulated noise, conditions $N=\{1,2,3,4\}$ differed from each other and all other conditions apart from the pairs (2, 3) and (4, 8). No conditions with $N>6$ differed significantly from speech-shaped noise, although the $N=\{6,8\}$ conditions differed from $N=\{64,128\}$.

IV. DISCUSSION

The experiment demonstrated that the masking effectiveness of N-talker babble varies nonmonotonically with N. This confirms the findings reported in the abstract of Carhart *et al.* (1975) and extends to speech-shaped noise the results of Danhauer and Leppler (1979). Masking by babble-modulated noise increased monotonically up to $N=6$ then leveled out. This confirms and extends to larger N the results of Bronkhorst and Plomp (1992) who measured the SRT of babble-modulated noise for $N=\{1,2,4,6\}$. The difference between the babble-modulated noise and natural babble conditions also varied with N.

An unexpected outcome of the study was the finding that all babble noises consisting of between 8 and 128 talkers have approximately the same masking effectiveness. The difference between the babble-modulated noise and natural babble conditions is usually attributed to the perceptual masking (Carhart *et al.*, 1969) which occurs when portions of the masker are wrongly attributed to the target speech. Phonetic cues are audible in the masker for small N, but become progressively inaudible as N increases. It is difficult to see how the factors which govern overall masking (energetic and informational) at $N=8$ could be the same as those which limit performance at $N=128$. Several studies have suggested that what is presumably the "linguistic uncertainty" component of informational masking effects is most potent for $N=2$ (Freyman *et al.*, 2004) or $N=3$ (Carhart *et al.*, 1975), and is almost absent by $N=10$ (Freyman *et al.*, 2004). In fact, the effect of informational masking may be underestimated because, for low values of N, listeners may be able to use level differences between the target and individual talkers in the background to help overcome some of the effects of linguistic confusions (Brungart, 2001).

Several factors might contribute to the breadth of the dip between $N=8$ and 128. Babble is identifiable as a signal composed of multiple speech sources for this range of N. Consequently, it is possible that attentional resources are devoted to monitoring the background in case some important speech event emerges. A related factor is the auditory system's response to fluctuating stimuli and, in particular, the enhancement of onsets. As N increases, the number of onsets in the background will increase, perhaps distracting attention from the target speech. Forward masking may also increase with N. However, the difference in masking effectiveness between natural babble and babble-modulated noise remains significant for large N, so any effects of distracting onsets and forward masking must be greater in the natural babble background. Another possibility is that the auditory system, like most systems for robust automatic speech recognition, makes use of background noise estimates to improve identification of the target. As the background becomes more stationary, the accuracy of the estimate increases. Ainsworth and Meyer (1994) found that the identification of syllables in steady-state noise was better when the noise was continuously present than when it was gated with the syllables. It is possible that, even though gated presentation was used in the current study, repeated exposure to the noise in the presence of a static /ɑ/ context is sufficient to provide a better noise estimate for $N>128$ than for smaller values of N. However, if the slight reduction in masking observed in babble-modulated noise for $N>64$ reflects increasing masker stationarity, then its effect is small.

These findings have yet to be generalized to other SNRs and different speech material, although a similar pattern of results was observed in a pilot study at SNRs of 0 and −12 dB. For the task and SNR condition investigated in this study (consonant identification in a VCV context at a SNR of −6 dB) an eight-talker babble provided the greatest amount of masking. Maximal informational masking for sentence material is usually reported to occur for small values of N (e.g., Freyman *et al.*, 2004). For consonants embedded in a static vowel context, listeners are likely to focus on brief acoustic cues present in the central region of the VCV in the absence of the wider contextual cues present in words and sentences. Conflicting cues to brief acoustic events are certainly salient in eight-talker babble and are perhaps sufficiently numerous in this condition to be at their most disruptive.

The N-babble continuum presents a challenge for accounts of speech perception in noise. The observed pattern of results appears to represent the combined contribution of several factors, each of which vary with N. Energetic masking increases with N, while linguistic masking reaches a peak at small values of N. Attentional demands resulting from monitoring a speechlike background, the distracting effects of numerous onsets and nonstationarity may continue to have a role for larger values of *N*.

ACKNOWLEDGMENTS

The authors would like to thank Jon Barker for discussion on an earlier draft of this letter. They also thank Jose Benki, one other reviewer, and the editor for their useful comments on the manuscript.

Ainsworth, W. A. and Meyer, G. F. (**1994**). “Recognition of plosive syllables in noise: Comparison of an auditory model with human performance,” J. Acoust. Soc. Am. **96**, 687–694.

Assmann, P. and Summerfield, Q. (**2004**). “The perception of speech under adverse acoustic conditions,” in *Speech Processing in the Auditory System*, Springer Handbook of Auditory Research Vol. 18, edited by S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay (Springer, Berlin).

Bronkhorst, A. W. (**2000**). “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,” Acust. Acta Acust. **86**, 117–128.

Bronkhorst, A. W. and Plomp, R. (**1992**). “Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing,” J. Acoust. Soc. Am. **92**, 3132–3139.

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

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). “Informational and energetic masking effects in the perception of multiple simultaneous talkers,” J. Acoust. Soc. Am. **110**, 2527–2538.

Carhart, R., Johnson, C., and Goodman, J. (**1975**). “Perceptual masking of spondees by combinations of talkers,” J. Acoust. Soc. Am. **58**, 535.

Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). “Perceptual masking in multiple sound backgrounds,” J. Acoust. Soc. Am. **45**, 694–703.

Cutler, A., Weber, A., Smits, R., and Cooper, N. (**2004**). “Patterns of English phoneme confusions by native and non-native listeners,” J. Acoust. Soc. Am. **116**, 3668–3678.

Danhauer, J. L. and Leppler, J. G. (**1979**). “Effects of four noise competitors on the California Consonant Test.” J. Speech Hear Disord. **44**, 354–362.

Duquesnoy, A. J. (**1983**). “Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons,” J. Acoust. Soc. Am. **74**, 739–743.

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

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.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (**1992**). “DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM,” NIST, Md.

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (**1977**). “Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability,” J. Acoust. Soc. Am. **61**, 1337–1351.

Markham, D. and Hazan, V. (**2004**). “The effect of talker- and listener-related factors on intelligibility for a real-word, open-set perception test,” J. Speech Lang. Hear. Res. **47**, 725-737.

Miller, G. A. (**1947**). “The masking of speech,” Psychol. Bull. **44**, 105–129.

Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., and Wang, X. (**1999**). “Consonant recordings for speech testing,” J. Acoust. Soc. Am. **106**, L71–L74.

Snell, K. B., Mapes, F. M., Hickman, E. D., and Frisina, D. R. (**2002**). “Word recognition in competing babble and the effects of age, temporal processing, and absolute sensitivity,” J. Acoust. Soc. Am. **112**, 720–727.