

Speech production modifications produced by competing talkers, babble, and stationary noise

Youyi Lu^{a)} and Martin Cooke

Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, United Kingdom

(Received 2 June 2007; revised 22 July 2008; accepted 27 August 2008)

Noise has an effect on speech production. Stationary noise and babble have been used in the past but the effect of a competing talker, which might be expected to cause different types of disruption, has rarely been investigated. The current study examined the acoustic and phonetic consequences of N -talker noise on sentence production for a range of values of N from 1 (competing talker) to infinity (speech-shaped noise). The effect of noise on speech production increased with both the number of background talkers (N) and noise level, both of which act to increase the energetic masking effect of the noise. In a background of stationary noise, noise-induced speech was always more intelligible than speech produced in quiet, and the gain in intelligibility increased with N and noise level, suggesting that talkers modify their productions to ameliorate energetic masking at the ears of the listener. When presented in a competing talker background, speech induced by a competing talker was more intelligible than speech produced in quiet, but the scale of the effect was compatible with the energetic masking effect of the competing talker. No evidence was found of modifications to speech production which exploited the temporal structure of a competing talker.

© 2008 Acoustical Society of America. [DOI: 10.1121/1.2990705]

PACS number(s): 43.72.Ar, 43.71.Gv, 43.72.Dv, 43.70.Mn [DOS]

Pages: 3261–3275

I. INTRODUCTION

Speech communication frequently takes place in the presence of background noise, conditions known to lead to changes in speech production, collectively known as the Lombard effect (Lombard, 1911). The goal of noise-induced modifications to normal speech production is not yet clear. It has been suggested that by modifying their vocal effort, speakers attempt to maintain a constant level of intelligibility in the face of degradation of the message by the environmental noise source (Summers *et al.*, 1988) and indeed some studies have reported intelligibility gains for “Lombard speech” presented in noise when compared to normal speech in noise (Dreher and O’Neill, 1957; Pittman and Wiley, 2001). However, the issue of how noise-induced speech production changes lead to the intelligibility gain has not yet been addressed.

Many studies have examined the acoustic-phonetic consequences of background noise on speech production (Hanley and Steer, 1949; Dreher and O’Neill, 1957; Charlip and Burk, 1969; Pisoni *et al.*, 1985; Stanton *et al.*, 1988; Summers *et al.*, 1988; Bond *et al.*, 1989; Junqua, 1993, 1996; Letowski *et al.*, 1993; Tartter *et al.*, 1993; Steeneken and Hansen, 1999; Garnier *et al.*, 2006; Pittman and Wiley, 2001; Varadarajan and Hansen, 2006). These studies have converged on a set of primary acoustic changes seen in Lombard speech relative to speech produced in quiet conditions. Specifically, Lombard speech demonstrates an increase in fundamental frequency (F0), speech level, and vowel duration, as well as a flattening of spectral tilt (more energy at higher

frequencies). The first and second formant frequencies (F1 and F2) also shift, with the consensus that F1 tends to increase while F2 has been reported to increase (Junqua, 1993) or decrease (Pisoni *et al.*, 1985). In addition, in Lombard speech, energy shifts between different classes of phoneme have been found. Junqua (1993) and Womack and Hansen (1996) reported a shift of energy from consonant to vowel while Hansen (1996) observed energy shifts from semivowel to vowel and consonant. Furthermore, improvements in automatic speech recognition performance under noisy conditions have been reported when Lombard effects have been incorporated into the recognizer (Hansen and Bria, 1990; Hansen, 1994; Chi and Oh, 1996; Hansen, 1996).

Many factors can influence the size of the acoustic changes observed in noise-induced speech modifications. Dreher and O’Neill (1957) reported that increasing the level of the masking noise from 70 to 100 dB resulted in a steady increase in duration from 15% to 31%, and a 6 to 9 dB increase in intensity, over that in quiet. In addition, noise level affects the scale of changes to fundamental frequency, spectral tilt, and formant frequencies (Summers *et al.*, 1988; Tartter *et al.*, 1993). The spectral tilt of the background noise has also been found to influence the Lombard effect. Junqua *et al.* (1998) reported that duration and fundamental frequency tend to increase with noise spectral tilt. Other factors affecting the size of the Lombard effect include the role of the word in a sentence, language spoken, and speaker gender. Patel and Schell (2008) observed larger effects of F0 and duration for information-bearing word types. The effect size was also larger for American English than French (Junqua, 1996). Junqua (1993) reported that the influence of the Lombard effect on vocal effort and F0 was greater for male speakers than for females. In addition, significant inters-

^{a)}Electronic mail: y.lu@dcs.shef.ac.uk

peaker differences exist in the range of production modifications seen (Stanton *et al.*, 1988; Junqua, 1993).

Acoustic differences between speech produced in quiet and in noise lead to differences in intelligibility. Dreher and O'Neill (1957), Pittman and Wiley (2001), and Summers *et al.* (1988) reported that for the same signal-to-noise ratio (SNR) with isolated words or continuous speech, speech produced in noise is more intelligible than speech produced in quiet. Dreher and O'Neill (1957) suggested that the changes in the spectral and temporal properties of speech which accompany the Lombard effect lead to an improvement in speech intelligibility. Summers *et al.* (1988) also reported that differences in the acoustic-phonetic structure of utterances produced in noise resulted in consistent increases in intelligibility across SNRs and talkers (although only two talkers were used). The magnitude of these effects increased as the environment became more severe. The influence on intelligibility of changes in acoustic parameters such as word length, vocal effort, and consonant-to-vowel energy ratio has also been studied. Howes (1957) showed that intelligibility increases with word length. Pickett (1956) reported that the intelligibility of speech increased with increase in vocal effort when the speech level was low (below 55 dB), remained constant over the range 55–78 dB, but dropped with increasingly forceful shouting (above 78 dB).

One aspect of noise-induced speech production changes which has received little attention is the effect of masking noise with different number of background talkers. Most studies have used stationary noise, though some have employed multitalker babble (Junqua, 1994; Pittman and Wiley, 2001; Garnier *et al.*, 2006). Junqua (1994) discovered that multitalker babble noise led to a larger vowel duration increase as compared to white-Gaussian noise. Garnier *et al.* (2006) demonstrated that increases in voice intensity, spectral energy, and word duration were greater in white noise than in cocktail party noise while mean F0 increased more in cocktail party noise than in white noise. However, wideband noise and multitalker babble did not appear to differentially influence the production of speech (Pittman and Wiley, 2001).

Surprisingly, the effect of an independent single-competing talker on speech production has not been investigated in depth.¹ In this regard, the study of Webster and Klumpp (1962) is relevant. In their study, talker-listener pairs were seated face to face and communicated word lists in conditions of quiet and ambient noise. When there was one background talker-listener pair, the speech level of the foreground talker increased by up to 9 dB, compared to the condition without the background pair. The speaking rate in words per second decreased slightly when the background pair was present. It was also found that the foreground pair made more communication errors when talking at the same time as the competing pair.

In speech perception studies, it is known that a competing talker generates masking effects which differ in two ways from stationary noise. First, at any given global SNR, speech is a far less effective energetic masker than stationary noise (Festen and Plomp, 1990; Simpson and Cooke, 2005). Energetic masking (EM) may be defined as that which results at

the auditory periphery due to overlapping excitation patterns. Second, speech on speech produces additional perceptual masking (Carhart *et al.*, 1969) over and above that caused by purely energetic factors, and indeed, this form of “informational masking” (IM) is the dominant effect in determining the intelligibility of a speech target masked by a competing utterance (Brungart, 2001). IM refers to all nonperipheral causes of elevations in masked threshold and includes incorrect assignment of elements in the acoustic mixture to the target speech as well as higher-level factors such as competition for limited attentional resources.

Since speech and noise maskers differ in the degree of EM and IM they produce in speech perception, it is of interest to discover whether they have differing effects on speech production. While the task of speech production in noise differs from speech perception in noise, production might be influenced by perceptual concerns in a number of ways. First, masking noise renders monitoring of a speaker's own productions more difficult, both energetically via loss of information of potential use in feedback and, informationally, due to competing attention. Second, speakers may be able to predict the masking effect of noise in the communicative environment at the ears of their interlocutor. In both cases, alterations to normal speech production might be expected.

The primary purpose of the current study was to determine how noise-induced speech production changes are affected by the degree of EM and IM potential of the noise. To measure the effect of differing amounts of EM and IM, N -speaker babble noises were employed for a range of values of N including $N=1$ (single speaker) and $N=\infty$ (speech-shaped noise). While EM increases with increasing N (Bronkhorst and Plomp, 1992; Simpson and Cooke, 2005), the influence of IM for sentence material is strongest for small N (e.g., $N=2$, Freyman *et al.*, 2004; $N=3$, Carhart *et al.*, 1975) and for $N=8$ for vowel-consonant-vowel tokens (Simpson and Cooke, 2005). Consequently, a number of intermediate values of N were also used in this study, and, in particular, we were interested in the effect of varying N on utterance-level properties such as duration, intensity, and fundamental frequency as well as formant frequencies, energies, and spectral energy distribution at the phonemic level. A further aim was to investigate whether talkers could exploit temporal fluctuations in the noise which are particularly profound for small values of N .

The intelligibility of noise-induced speech is known to increase over speech produced in quiet when noise is added (Dreher and O'Neill, 1957; Summers *et al.*, 1988; Junqua, 1993). A secondary goal of the study was to measure speech intelligibility as a function of the number of talkers and level of background noise. There is still no clear idea of the origin of these intelligibility gains. The current study employed a computational model of EM in an attempt to determine whether the acoustic changes produced by noise-induced speech result from an attempt to reduce the EM effect at the listener's ears.

II. SPEECH PRODUCTION IN NOISE

A. Corpus design

To determine how speech production changes in the presence of widely differing maskers, a corpus of N -talker babble maskers for $N=\{1,2,4,8,16,\infty\}$ was produced. These values were chosen based on an earlier study which measured the masking effect of N -talker babble for a large number of values of N (Simpson and Cooke, 2005). Since one goal of the study was to investigate the role of IM on speech production, talkers produced sentences which were similar in form to those used to produce N -babble maskers. The Grid corpus (Cooke *et al.*, 2006) was used as the source of the masking material, and talkers were asked to read sentences from this corpus. Grid consists of simple six-word sentences such as “lay green with A4 now” or “set white at B8 again.” Grid has been used in speech-on-speech tasks and shown to produce large amounts of IM (Cooke *et al.*, 2008), and the noise-intelligibility relation for speech-shaped noise has been measured (Barker and Cooke, 2007). Maskers for the six values of N were presented at 89 dB sound pressure level (SPL), a level in the middle of the range known to induce significant speech production changes (Stanton *et al.*, 1988 used 90 dB; Summers *et al.*, 1988 used 80, 90, and 100 dB; Junqua, 1993 used 85 dB). To examine the effect of noise level for the extreme values of N , the single speaker ($N=1$) and speech-shaped noise ($N=\infty$) maskers were also presented at 82 and 96 dB SPL. Finally, a “quiet” condition was used to provide a reference against which noise-induced speech production modifications could be measured. In summary, talkers produced speech in a total of 11 conditions ($6 \times N$ values at 89 dB SPL, $2 \times N$ values at 82 dB SPL, $2 \times N$ values at 96 dB SPL, and quiet). Symbols used to represent the ten noise conditions here and elsewhere are in the form $N(\text{number of talkers})_{(\text{level})}$ so that, for example, $N1_{89}$ refers to a competing talker background at 89 dB level, while $Ninf_{96}$ indicates a speech-shaped noise background at 96 dB level.

B. Sentence lists and maskers

To allow comparison of acoustic and acoustic-phonetic properties, talkers produced the same set of 50 sentences conforming to the Grid syntax in each of the 11 conditions. However, to introduce some variation, each talker produced a different set of 50 sentences. N -babble maskers for the finite values of N were generated by adding utterances drawn at random from the Grid corpus into a 60 s circular buffer until the required babble density was obtained. This approach avoids problems with uneven masking effects which would have occurred if utterances had been added with synchronized start times. One consequence of this strategy was that masking sentences were not synchronized with the talker’s productions: background utterances would start at a random point in the sentence and there could be a change in talker during the time allotted to the production of a single utterance. Prior to incorporation into the buffer, leading and trailing silence was removed, and utterances were scaled to have equal root mean square (rms) levels. Masking noise to accompany individual talker productions consisted of 3 s

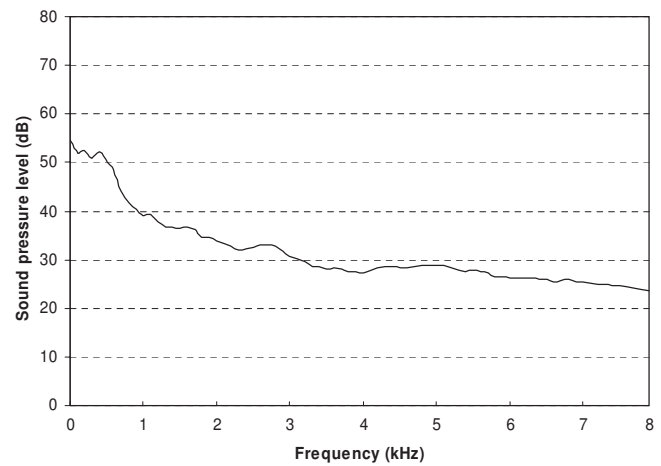


FIG. 1. Long-term average speech spectrum for the Grid corpus.

segments of babble drawn at random from the 60 s buffer. Speech-shaped noise was produced by filtering white noise with a filter whose spectrum equalled the long-term spectrum of the Grid corpus, as shown in Fig. 1. Again, a 60 s segment was generated for subsequent random selection.

C. Talkers

Eight native speakers of British English (four males and four females) drawn from staff and students in the Department of Computer Science at the University of Sheffield participated in the corpus collection. All received a hearing test using a calibrated software audiometer which was used to test each ear separately at the six frequencies: 250, 500, 1000, 2000, 4000, and 8000 Hz. One participant had a slight hearing loss (23 dB hearing level (HL)) in one ear at the highest frequency (8 kHz) but was retained for the study. The remaining participants had normal hearing (better than 20 dB hearing level in the range of 250–8000 Hz). Ethics permission was obtained following the University of Sheffield Ethics Procedure. Talkers were paid for their participation.

D. Procedure

Corpus collection sessions took place in an IAC single-walled acoustically isolated booth. Speech material was collected using a Bruel & Kjaer (B & K) type 4190 $\frac{1}{2}$ in. microphone coupled with a preamplifier (B&K type 2669) placed 30 cm in front of the talker. The signal was further processed by a conditioning amplifier (B & K Nexus model 2690) prior to digitization at 25 kHz with a Tucker-Davis Technologies (TDT) System 3 RP2.1. Simultaneously,² maskers were presented diotically over Sennheiser HD 250 Linear II headphones using the same TDT system. Speakers wore the headphones throughout, including for the quiet condition. Of course, the use of closed headphones to deliver masking noise can be expected to introduce frequency-dependent own-voice attenuation (Arlinger, 1986; Bofill *et al.*, 2006). Since the current study involves comparison across masking conditions, the constant attenuation characteristics were not considered to be an important factor. However, the closed headphone setup was compared with a com-

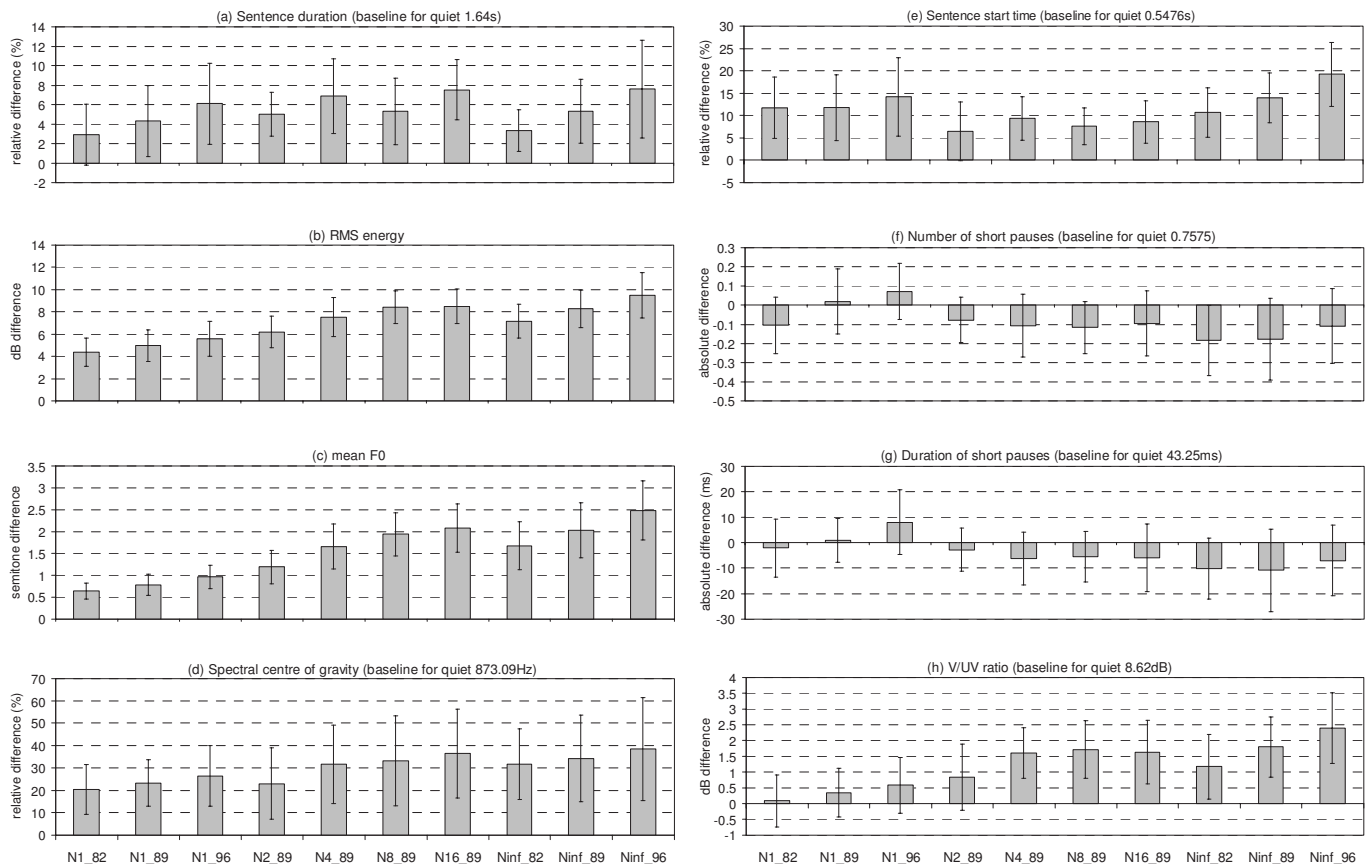


FIG. 2. Differences between acoustic parameter values for each noise condition compared to speech produced in quiet. Where meaningful, “baseline” parameter values in quiet are given in order to provide an absolute reference. Values shown are means over talkers and error bars, here and elsewhere, indicate 95% confidence intervals. Noise conditions are indicated as $N(\text{number of talkers})_{\text{level}}$.

compensated transmission channel for a subset of conditions. The chief finding was that the recording method was not a significant factor in the speech production modifications measured (see Appendix for details).

Sentence collection and masker presentation was under computer control. Talkers were asked to read out sentences presented on a computer screen and had 3 s to produce each sentence and were allowed to repeat the sentence if they felt it necessary. All the repetitions were saved to allow analysis of the number of “false starts” in the different masking conditions. Prior to saving, signals were scaled to produce a maximum absolute value of unity to make best use of the amplitude quantization range. Scale factors were stored to allow the normalization process to be reversed.

Talkers recorded the 11 conditions over two sessions of 30 min each on two days. They were familiarized with the type of sentences and the task before each collecting session. The three single-talker conditions were combined into a single block as were the three speech-shaped noise conditions, and both sentence order and masker level were randomized within the block. Thus, the 11 conditions were presented in seven blocks, and block order was randomized for each talker.

E. Postprocessing

In order to measure acoustic parameters at the level of individual phonemes, a set of speaker-independent phoneme-

level hidden Markov models (HMMs) was built from speech material in the Grid corpus (Cooke *et al.*, 2006) using the HTK HMM toolkit (Young *et al.*, 1999). These models were used to produce phoneme-level transcriptions of the collected utterances via forced alignment using the HVITE tool in HTK. Leading and trailing silent intervals identified via the alignment process were removed. For each talker in each of the 11 conditions, transcriptions of a random selection of 10% of the utterances were manually inspected and found to be accurate.

III. ACOUSTIC AND ACOUSTIC-PHONETIC ANALYSES

A. Utterance-level analysis

Eight acoustic properties were estimated for each utterance. Sentence duration, rms energy, mean fundamental frequency (F0), and spectral center of gravity (CoG) were computed via PRAAT 4.3.24 (Boersma and Weenink, 2005).³ Sentence start time (i.e., the onset of speech production relative to the onset of the interfering signal) and the number and duration of short pauses (>20 ms) were computed using phoneme-level transcriptions. These latter measures were motivated by the possibility that talkers might avoid overlapping with the background signal, especially in the competing speech conditions. Finally, the voiced-to-unvoiced energy ratio (V/UV ratio) was estimated.

Differences between across-talker means in each background compared to the quiet condition are shown in Fig. 2

TABLE I. Summary of the results of statistical analyses comparing the values of acoustic parameters for speech produced in quiet with speech produced in noise. Column N_{89} represents six N -talker conditions ($N=\{1, 2, 4, 8, 16, \infty\}$) at a noise level of 89 dB. Columns $N1$ and $Ninf$ represent three level conditions (level= $\{82, 89, 96$ dB}) for $N=1$ and $N=\infty$, respectively. The final ten columns represent the individual noise conditions as follows. (1–3) $N=1$, levels 82, 89, and 96 dB; (4–7) $N=\{2, 4, 8, 16\}$ at 89 dB; (8–10) $N=\infty$, levels 82, 89, and 96 dB. Symbols “ \uparrow ” and “ \downarrow ” represent significant increases or decreases in the parameter in noise over the quiet condition. Significance levels: *** <0.001 , ** <0.01 , and * <0.05 .

	Repeated-measures ANOVA			One-sample t -test (test value=0)									
	N_{89}	$N1$	$Ninf$	1	2	3	4	5	6	7	8	9	10
Sentence duration		* \uparrow	* \uparrow			* \uparrow	** \uparrow	** \uparrow	* \uparrow	** \uparrow	* \uparrow	* \uparrow	* \uparrow
rms energy	*** \uparrow	*** \uparrow	** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow
Mean F0	*** \uparrow	** \uparrow	** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	*** \uparrow	** \uparrow	*** \uparrow	*** \uparrow
CoG	* \uparrow	* \uparrow		** \uparrow	** \uparrow	** \uparrow	* \uparrow	* \uparrow	* \uparrow	** \uparrow	** \uparrow	* \uparrow	* \uparrow
Sentence start time			** \uparrow	* \uparrow	* \uparrow	* \uparrow		** \uparrow	** \uparrow	** \uparrow	** \uparrow	** \uparrow	** \uparrow
No. of short pauses	* \uparrow	** \uparrow											
Duration of short pauses		* \uparrow											
V/UV ratio	* \uparrow		* \uparrow					*** \uparrow	*** \uparrow	*** \uparrow	** \uparrow	*** \uparrow	*** \uparrow

for each of the eight acoustic parameters. The number of talkers and noise level in each background is shown as is the baseline mean for the parameter (that is, the mean value in the quiet condition).

To aid the interpretation of Fig. 2, several statistical analyses were carried out for each acoustic parameter. A repeated-measures analysis of variance (ANOVA) analyzed the effect of the number of talkers (N) in the maskers at the 89 dB level. To determine any interaction effect between N and noise level, a two-way repeated-measures ANOVA with within-subjects factors of N (1, ∞) and masker level (82, 89, 96 dB) was computed. Two further single-factor repeated-measures ANOVAs examined the effects of noise level in the single talker and speech-shaped noise condition. Finally, one-sample t -tests (test value=0) were employed to determine the significance of differences between each masking condition and quiet.

Table I summarizes the results of the statistical analysis for each of the utterance-level acoustic measurements. Many parameters demonstrated significant increases in most of the noise backgrounds compared to quiet (final ten columns of Table I). The most significant effects were for energy (which increased by between 3 and 9 dB relative to quiet) and mean F0 (0.6–2.5 semitones). Spectral CoG increased from the quiet baseline of 870 Hz by 20%–38%. The mean sentence duration in quiet of 1.64 s rose by 2.4%–7.6%, while the pause before speaking increased by 6%–18% from a baseline of 0.55 s in quiet. The V/UV ratio rose in most conditions, from 8.6 dB in quiet by up to 2.4 dB. No significant overall effect of the duration of short pauses or the number of short pauses was found.

For some parameters, the difference between speech produced in quiet and that produced in the presence of noise increased with the number of talkers in the babble (column 2 of Table I). The strongest effect of N was seen for energy and F0, with a lesser effect of spectral CoG and V/UV ratio. However, the effect of N typically reached a plateau at around $N=8$ talkers. For duration, energy, and mean F0, the effect of noise level was similar for the single talker and speech-shaped noise backgrounds (columns 3 and 4 of Table I). Sentence start time (the delay before the talker started

speaking following the onset of the noise) increased with level for the speech-shaped noise condition, while the number (and to a lesser extent, duration) of short pauses increased with level in the single-competing talker condition. Similarly, effects of CoG and duration of short pauses were only found in the competing talker conditions, while V/UV ratio increased in the stationary noise conditions. No interaction between the effects of N and noise level was found for any of the parameters.

Figure 2 also indicates the range of talker variation for each parameter and background. While all talkers showed similar changes in energy and F0, significant cross-talker variability is present for the remaining measures.

Increases in duration, mean energy, and F0 in stationary noise were also found in previous studies on Lombard speech using words and short sentences (Dreher and O’Neill, 1957; Pisoni *et al.*, 1985; Summers *et al.*, 1988; Bond *et al.*, 1989; Letowski *et al.*, 1993; Tartter *et al.*, 1993; Steeneken and Hansen, 1999; Garnier *et al.*, 2006). The V/UV ratio increased in most of the N -talker conditions, echoing the findings of Junqua (1993) and Womack and Hansen (1996) for stationary noise. However, a pattern of reduced sentence duration was found by Varadarajan and Hansen (2006), who also reported a decrease in short pause duration, while no such effect was found here. Varadarajan and Hansen (2006) suggested that decreases in sentence and short pause duration could be caused by a sense of urgency on the part of the speaker, which occurred due to the constant exposure to the background noise. Here, the fact that noise was presented only when the talker was due to speak might account for the differences.

B. Phoneme-level analysis

Prior to phoneme-level analysis, all of the postprocessed utterances were normalized to have equal rms energy. Individual phonemes of the utterances were segmented via the phoneme-level transcriptions. Phonemes were grouped into the six categories {vowel, diphthong, liquid, fricative, plo-

TABLE II. Phoneme categories.

Vowel	i:, ɪ, e, u:, æ
Diphthong	eɪ, aɪ, aʊ,
Liquid	w, l, r
Fricative	f, s, v, z, ð
Plosive	p, t, k, b, d, g
Nasal	n

sive, nasal}, as shown in Table II. Phonemes which had fewer than 500 instances were not used. On average, around 3000 instances of each phoneme were employed.

Duration was measured for all phoneme instances, while spectral CoG was measured for all apart from the plosives, whose more complex spectrotemporal development precluded a meaningful measurement. Spectral tilt was computed for all the vowels. It is important to note that due to the limited number of contexts present in the Grid corpus, the phoneme instances used in this analysis should not be regarded as prototypical. For instance, in Grid, the /æ/ vowel can only be found in the word “at,” most of which were reduced to schwa in this context. As a consequence, a formant analysis (frequencies, energies, and bandwidths) was undertaken solely for the vowels /i:/, /ɪ/, /e/, and /u:/ in the words “green,” “bin,” “red,” and “soon,” respectively. Frequency and energy values were computed as the average of the central three frames in each vowel instance. All of the measurements apart from spectral tilt were computed using the Praat program V 4.3.24 (Boersma and Weenink, 2005). For spectral tilt, the spectrum of an entire phoneme instance was divided into ten energy bands following Stanton *et al.* (1988). Spectral tilt was estimated as the slope of the best linear fit to the ten log energy values. Individual talker and overall measurements were computed for each phoneme. Measurements were obtained by averaging the differences between the phoneme instances of the utterances from each of the ten *N*-talker conditions and the instances in the same position of the same speech sentences from the quiet condition. Individual and overall measurements for all the acoustic properties are expressed as relative percentage differences from quiet, apart from formant frequency and bandwidth, which were expressed as Hertz difference, and energy, which used difference in decibels.

Figures 3 and 4 display the quantitative results of the phoneme-level analysis. To enhance the readability of the plots, results have been averaged across subsets of the ten noise backgrounds. In general, changes in noise backgrounds over quiet were found, and stronger effects were observed for larger number of background talkers and for higher noise levels.

Compared to quiet, increases in *N* and masker level led to an increase in the duration of most sound types apart from the fricative /f/ and the nonalveolar plosives, for which a slight shortening was observed. Increases in spectral CoG were seen for all sounds. For most, the increase was substantially larger than 25%, although the fricatives /f/ and /s/ showed only modest increases. Similar findings for the duration and CoG of vowels have been reported (Junqua, 1993; Stanton *et al.*, 1988; Garnier *et al.*, 2006). Vowel spectral tilt

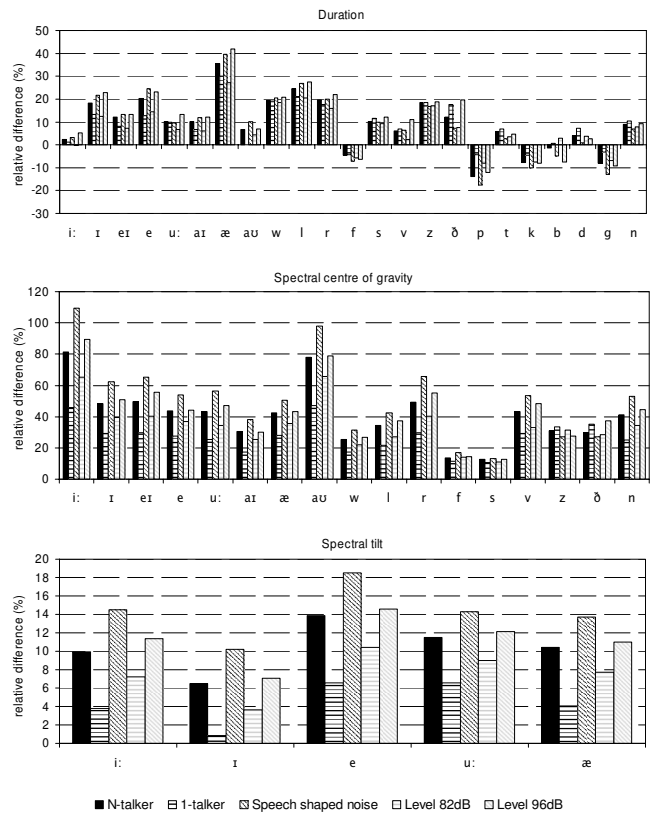


FIG. 3. Phoneme-specific differences in duration (top), spectral CoG (middle), and spectral tilt (bottom) in noise and quiet conditions. For ease of display, the noise conditions are grouped into five subsets. *N*-talker: all ten noise backgrounds; one talker: *N*_{1_82}, *N*_{1_89}, and *N*_{1_96}; speech-shaped noise: *N*_{inf_82}, *N*_{inf_89}, and *N*_{inf_96}; level 82 dB: *N*_{1_82} and *N*_{inf_82}; level 96 dB: *N*_{1_96} and *N*_{inf_96}.

became flatter in all conditions, with differences in degree between the vowels. Other studies have reported a similar pattern (Pisoni *et al.*, 1985; Summers *et al.*, 1988; Varadara-jan and Hansen, 2006).

In addition, similar statistical analyses to those used for utterance-level parameters in Sec. III A were carried out for formant frequencies, energies, and bandwidths for each vowel. For speech produced in noise, F1 frequency increased significantly by up to 100 Hz. Such effects were stronger for speech-shaped noise, compared to a competing talker background, for all the vowels. F2 and F3 frequencies fell by as much as 60 and 80 Hz, respectively, but these tendencies were only statistically significant for the vowels /i:/ and /ɪ/. For F2 and F3 frequencies, no significant differences were found between competing talker and speech-shaped noise. Increases in vowel F1 frequency were also seen in earlier studies (Pisoni *et al.*, 1985; Summers *et al.*, 1988; Bond *et al.*, 1989; Junqua, 1993; Garnier *et al.*, 2006). For F2, Junqua (1993) reported increases for females while Pisoni *et al.* (1985) found the opposite for both males and females. Other studies (Summers *et al.*, 1988; Bond *et al.*, 1989; Garnier *et al.*, 2006) demonstrated a large amount of vowel and utterance-dependent pattern of F2 frequency change. Junqua (1993) suggested that the F3 frequency of vowels tends to remain constant in noise.

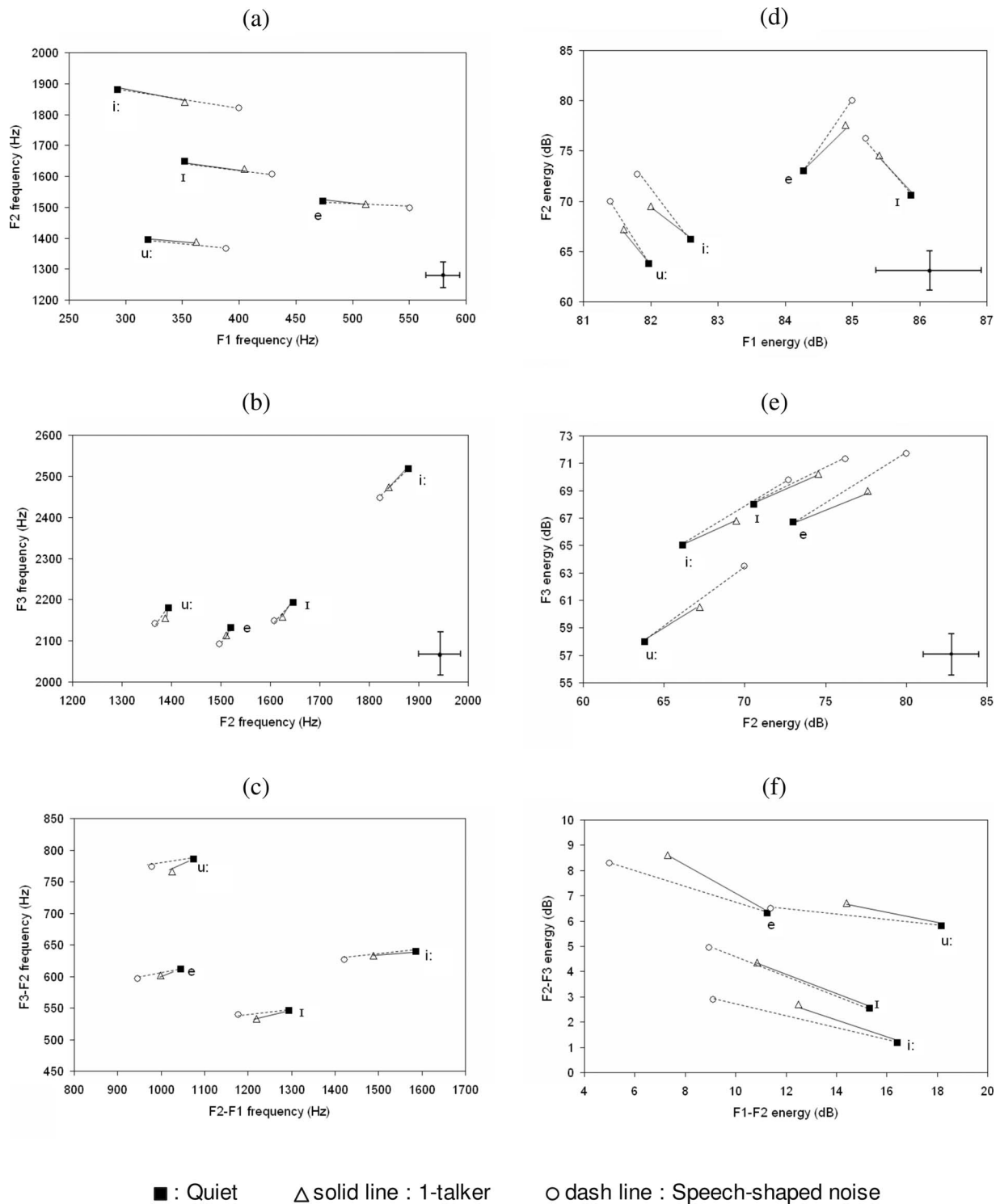


FIG. 4. Formant frequencies (left) and energies (right) for the vowels in “green,” “bin,” “red,” and “soon,” for speech produced in quiet and noise. In each case, values are averages taken from the central three frames over all instances of the vowels. For clarity, averages across the three single-talker and three speech-shaped noise conditions are shown. Error bars in lower-right corner indicate 95% confidence intervals.

Significant increases in F2 and F3 energy for the vowels compared to the quiet condition were measured. Such effects were significantly stronger for speech-shaped noise compared to the competing talker background. F1 energy changed little. The overall effect of formant energy changes is consistent with observed changes in spectral tilt.

Furthermore, significant increases in F1 bandwidth and decreases in F2 and F3 bandwidths for all the vowels

compared to the quiet condition were found. For most of the vowels, changes in F1 and F2 bandwidths tended to be significantly larger for speech-shaped noise compared to competing talker background while for F3 bandwidth, such tendencies were only significant for the vowels /i:/ and /ɪ/. The changes in F2 and F3 bandwidths for speech produced in noise are consistent with those reported in Hansen and Bria (1990). For F1 bandwidth, Hansen and Bria (1990) found an

increase for /i:/ and /ɪ/ and a decrease for /e/ while Junqua (1993) suggested a decreasing tendency for most of the vowels.

C. Correlation analysis

The above analyses treat speech production changes as independent of each other, but it is possible that correlated changes exist in acoustic parameters such as F0 and F1 frequencies as a result of speech energy changes. Correlations between energy and both F0 and F1 frequencies were investigated. The Pearson correlation coefficient between energy and F0 and energy and F1 frequency was computed independently for all voiced segments. To arrive at a single correlation measure, the weighted mean of segment-based correlations was derived, with weights given by segment duration. For energy versus F0, there was a slight but significant *decrease* ($p < 0.05$) in correlation in most of the noise conditions compared to quiet ($r = 0.37$). The correlation decreased significantly with an increasing number of background talkers [$F(2.9, 20.1) = 3.982$, $p < 0.05$]. For energy and F1 frequency, significantly *increased* ($p < 0.05$) correlation was found in all noise conditions compared to quiet ($r = 0.23$). Correlations also increased with the number of talkers [$F(3.8, 26.4) = 6.504$, $p < 0.01$].

D. Discussion

The current results generally confirm the effects of stationary noise on speech production found in previous studies, both at the level of overall acoustic parameter values and for individual phoneme classes. More importantly, they demonstrate for the first time the effect of the number of talkers making up the background babble, including the case of a single talker. For nearly all of the parameters where there is a significant difference between speech produced in stationary noise and in quiet, there is a similar, but smaller, effect when a single talker speaks in the background while speech is produced. Similarly, changes in noise level which have an effect in the stationary noise case tend also to affect the single-talker case. The effect of intermediate background conditions (i.e., multitalker babble for more than one talker) usually falls somewhere between the two extremes. For all parameters, no interaction between the effects of noise level and the number of background talkers was present. One interpretation of these results is that the Lombard effect is influenced by both noise level and number of background talkers, acting independently.

For those parameters which might be expected to reflect the differences in information conveyed by the background, namely, sentence start time and the statistics of short pauses, some small differences were found. There were more pauses longer than 20 ms in the single-talker background than in the other conditions. The pause prior to speaking was longer in the single-talker background than for most of the babble conditions, although the pause was slightly shorter than in the stationary noise case. It is possible that the noncommunicative task limited the scope for such effects.

Some acoustic effects might be the consequence of intentional changes while others may be secondary, caused by

articulatory constraints. For example, as pointed out by Gramming *et al.* (1988), the raising of subglottal pressure in order to create a louder voice causes an increase in F0. On the other hand, it is also possible that in the production of high-pitched voice, SPL is raised due to a larger number of speech pressure cycles per time unit resulted from the increase in F0. In addition, the wider jaw opening in order to increase sound amplitude induces an increase in the first formant frequency (Lindblom and Sundberg, 1971). In the current study, correlations between F0 and energy actually decreased in noise, although F1 frequency and energy became more correlated. Thus, it is possible that speakers were using intentional changes in both energy and F0 in response to noise. It is likely that other factors such as physiological and semantic constraints on possible F0/F1 values and range also limit the extent to which speakers can manipulate these parameters independently.

IV. INTELLIGIBILITY OF NOISE-INDUCED SPEECH

A. Motivation

Speech produced in the presence of noise can lead to increases in intelligibility over speech produced in quiet mixed with equivalent noise tokens at the same SNR (Dreher and O'Neill, 1957; Summers *et al.*, 1988; Junqua, 1993). The speech material collected in the current study employed a wider range of noise backgrounds, allowing several new issues to be explored. First, the general finding that the effect of noise on certain acoustic parameters tended to increase with both noise level and number of talkers (N) suggests that any intelligibility gains may also be influenced by noise level and N . Experiment I measured speech intelligibility as a function of noise level and N for noise-induced speech compared to speech produced in quiet with added noise.

When faced with the task of communicating in the presence of a single-competing talker, talkers might adopt strategies to reduce both the EM and IM components at the ear of the listener. Two further experiments explored these possibilities. In experiment II, listeners were presented with utterances masked by a competing talker. The intelligibility of utterances produced in quiet was measured and compared to that of the same set of utterances induced by a competing talker when presented in the background of the inducing competing talker maskers. Any intelligibility gains in the latter "matched" case might be interpreted as resulting from a talker's awareness of the IM effect of the competing utterance. However, increases in intelligibility could also be derived from reductions in EM due to acoustic changes in the competing speaker-induced utterances. Experiment III attempted to distinguish the two hypotheses by comparing the intelligibility of speech produced in the presence of a competing talker when presented in the matched competing talker background with the same utterances presented in an unmatched competing talker background. If talkers are sensitive to the IM potential of a specific competing utterance rather than the EM properties of speech in general, listeners should produce higher scores in the matched condition.

B. Experiment I: Sentences in stationary noise

1. Listeners

Twelve native speakers of British English (nine males and three females) drawn from the undergraduate and post-graduate population of the Department of Computer Science at the University of Sheffield took part in experiment I. All subjects received a hearing test using the same software and procedure, as described in Sec. II C. All had normal hearing apart from one participant with a hearing level of 25 dB in one ear at 8 kHz. This subject was retained for the study. Ethics permission was obtained following the University of Sheffield Ethics Procedure.

2. Stimuli

Utterances collected in quiet and in the presence of noise were presented in a background of stationary speech-shaped noise. Five sets of 100 utterances were used, corresponding to speech produced in quiet, in a background of a competing talker at levels of 82 and 96 dB SPL, and in a background of stationary noise at 82 and 96 dB SPL. In all five conditions of experiment I, utterances were mixed with a speech-shaped noise masker at an overall SNR of -9 dB, a value chosen on the basis of pilot tests to reduce ceiling and floor effects. Prior to mixing, target utterances were scaled to have the same rms level. Maskers were gated on and off with the endpointed utterances, and the mixed signals were scaled to a level of approximately 68 dB SPL.

3. Procedure

Experiment I took place in an IAC single-walled acoustically isolated booth. Stimuli presentation and results collection was controlled by a computer program. Stimuli were presented diotically over Sennheiser HD 250 Linear II headphones via a TDT System 3 RP2.1. Listeners were given instructions to identify in each noisy utterance the letter and digit keywords. This they did via a computer keyboard whose keys were selectively activated to minimize keying errors. For consistency with later experiments, in which the color keyword was used to identify the target utterance, sentences within each condition were organized into four blocks by color keyword. Condition order was balanced across listeners while both color blocks and utterance order within blocks were randomized for each listener. The experiment took place in a single session which was preceded by a short practice. In addition, four practice tokens were added to the start of each condition. Listeners were unaware of these tokens and they were not scored. The entire session required around 30 min to complete.

4. Results

For utterances produced in quiet and presented in speech-shaped noise, listeners obtained a mean keyword identification score of 42%. However, for the four conditions involving the identification of utterances produced in a noise background, keyword scores were substantially higher. As shown in Fig. 5, the increase in scores for noise-induced speech ranged from 9 to 25 percentage points. These increases were statistically significant ($p < 0.01$ in the single-

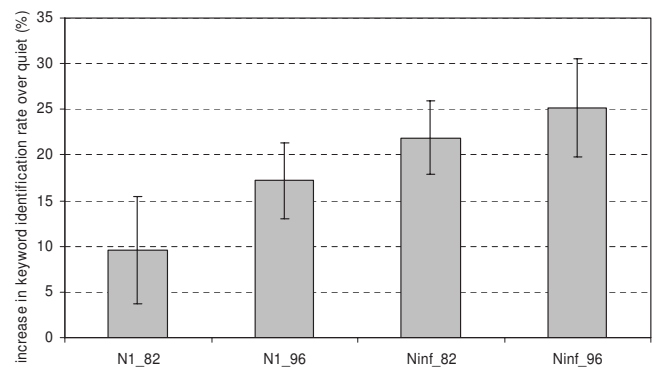


FIG. 5. Keyword identification rates for noise-induced speech over speech produced in quiet when added to speech-shaped noise (experiment I). The baseline keyword identification score for utterances produced in quiet is 42%.

talker 82 dB condition; $p < 0.001$ in the other three conditions). The two single-talker backgrounds led to the smallest improvements, and in both the single-talker and stationary noise backgrounds, the gain in intelligibility increased with noise level. Among the four noise-induced speech conditions, a two-way repeated-measures ANOVA with factors of $N = \{1, \infty\}$ and level = $\{82, 96$ dB] found a significant effect of N [$F(1, 11) = 27.276$, $p < 0.001$] and noise level [$F(1, 11) = 8.278$, $p < 0.05$]. The N by noise level interaction was not significant ($p > 0.2$).

C. Experiment II: Sentences in competing utterances

1. Listeners, stimuli, and presentation

Listeners who took part in experiment I also took part in this experiment. Four conditions tested the identification of keywords in utterances when presented in a competing speaker background. In two conditions, listeners heard speech produced in quiet conditions added to other speech material produced in quiet, drawn from the same corpus (Cooke *et al.*, 2006). In the other two conditions, listeners heard speech that was produced in a competing speech background added to that competing speech background. These “speech-induced” conditions were drawn from those collected as described in Sec. II and corresponded to the 82 and 96 dB background levels. Both “quiet” conditions were identical apart from the choice of sentences used for the background. Two conditions were used to enable the same set of speech maskers to be used in the speech-induced and quiet conditions.

As for experiment I, 100 utterances were used for each condition. For this experiment, sentences contained no keywords in common with those of the masker. Sentences were added so that the target to masker ratio was -9 dB, a value chosen on the basis of pilot experiments, and maskers were gated on and off with the endpointed target sentence. Due to the approach taken to the generation of competing speech maskers as described in Sec. II, the start of a sentence did not necessarily coincide with the start of a sentence in the masker. In this respect, the two-talker scheme was different from those used in IM experiments (e.g., Brungart, 2001; Cooke *et al.*, 2008).

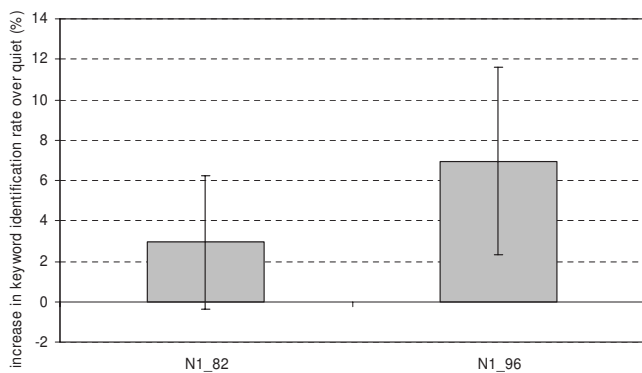


FIG. 6. Keyword identification rates for utterances induced by a speech background over utterances produced in quiet when added to the inducing speech (experiment II). The baseline keyword identification scores for utterances produced in quiet are 81% and 78%, respectively.

The stimulus presentation setup was as described in experiment I. Since this task involved identifying a target in a very similar masker, listeners required information to distinguish the target and masker sentences. The color keyword was used to indicate which utterances listeners had to attend to. The corpus contains four color keywords, so stimuli were organized into four blocks within each condition. At the start of each block, listeners were instructed (via the computer screen) to identify the letter and digit in the sentence containing a given color.

2. Results

Figure 6 displays the difference in keyword identification rates between the speech-induced and quiet utterances for the two levels 82 and 96 dB. While the speech-induced utterances are more intelligible for both levels, only the 96 dB case reaches statistical significance at the 0.05 level [$t(11)=1.756$], suggesting that speech produced in sufficiently intense backgrounds containing a single-competing talker is more intelligible than speech produced in quiet when added to the same competing talker material. This finding extends that of experiment I to a highly nonstationary masker. However, the absence of an effect for speech produced in less intense backgrounds calls into question the extent to which this effect is due to an attempt by the speaker to minimize the degree of IM at the ear of the listener.

D. Experiment III: Induced speech in matched and unmatched backgrounds

1. Listeners, stimuli, and presentation

Listeners who participated in experiments I and II also took part in this experiment. Experiment III compared two conditions, one in which the target material consisted of speech induced by other speech was presented in the background of the inducing speech material (“matched”) and one in which the same target speech was presented in “unmatched” backgrounds. Target speech consisted of utterances collected as described in Sec. II in the presence of a competing talker presented at 89 dB SPL. All other stimulus construction and presentation details were the same as for experiment II (100 utterances, -9 dB target-to-masker ratio,

and presentation of targets blocked by color keyword). Experiments II and III were performed in sequence in the same session, which lasted approximately 30 min.

2. Results

Keyword identification score in the matched condition was 1.4 percentage points higher than in the unmatched condition (score of 86%). However, this failed to reach statistical significance at the 95% level ($p=0.08$). This outcome suggests that, in this task, talkers do not modify their productions in response to the details of a specific competing utterance.

E. Discussion

The three perceptual experiments here explored the extent to which the presence of competing speech and stationary noise influences the intelligibility of speech productions. Experiment I confirmed previous findings on the increased intelligibility of speech produced in stationary noise backgrounds (Dreher and O’Neill, 1957; Summers *et al.*, 1988) and extended these results to single-talker maskers. The size of intelligibility gains was closely correlated with the extent of acoustic changes measured in Sec. III: stationary noise backgrounds and intense background level both resulted in larger intelligibility gains than single-talker backgrounds and less intense backgrounds. However, all backgrounds tested resulted in significant gains in intelligibility.

Experiments II and III employed maskers designed to invoke large amounts of IM to explore the possibility that talkers modify their production strategy dynamically in response to the presence of competing speech. Experiment II demonstrated that speech produced in an intense competing speech background was more intelligible than speech produced in quiet when presented in the same background. However, for speech produced in a less intense background, no such difference was found, suggesting that EM rather than IM is dominant since the lower background intensity during production (82 dB SPL) is still relatively strong and could be expected to produce IM effects. It seems likely that similar principles as those leading to modifications in production for speech produced in stationary noise backgrounds operate in the competing speech condition.

The results of experiment III do not support the idea that talkers modify their productions in response to the details of individual competing utterances in order to improve intelligibility at the ear of the listener. There was no significant difference in identification scores between speech produced in the speech backgrounds for the maskers which induced the utterances compared to the same induced utterances presented with random speech maskers.

V. DOES NOISE-INDUCED SPEECH OFFER MORE GLIMPSES OPPORTUNITIES?

A. Motivation

While the finding that noise-induced speech is often more intelligible when presented in noise has been reported in studies dating back many years (Dreher and O’Neill,

1957) and has been confirmed here, little effort has been directed toward an explanation of the intelligibility gain. Here, we test the hypothesis that the intelligibility of noise-induced speech is related to the availability of “glimpses” of speech at the ear of the listener in the presence of noise. Glimpses of a signal are defined as those connected regions in its spectrotemporal representation over a certain minimum “area” calculated from the number of spectrotemporal “pixels” and where each spectrotemporal pixel has a local SNR larger than a threshold (Cooke, 2006). This hypothesis is grounded in the EM produced by the masker and differs from a pure EM explanation in that glimpses incorporate the idea that listeners only have access to spectrotemporal regions which are sufficiently dominant in both local SNR and spectrotemporal extent to allow them to stand out above the masker (Cooke, 2006).

B. Glimpse measures

Two glimpsing statistics were measured for the signal mixtures used in the intelligibility experiments described in the previous section. One, “glimpse area,” is the number of spectrotemporal points where the glimpse criteria described above hold. Since glimpse area will typically increase with signal duration, “glimpse proportion” was also computed, defined as the proportion of spectrotemporal points which meets the glimpse criteria. This latter measure is independent of duration and helps to distinguish simple speech production processes which improve glimpsing opportunities by slowing speech rate from those which reallocate energy in time and frequency to improve glimpsing opportunities.

Computation of glimpse measures was based on a spectrotemporal excitation pattern (STEP) representation formed for the target and masker independently. A STEP is produced by first passing the time-domain signal through a 64 channel gammatone filterbank, smoothing the Hilbert envelopes, integrating the energy into 10 ms frames, followed by log compression. More details of the computation can be found in Cooke (2006). Following Cooke (2006), a minimum area of 5 and a local SNR of -5 dB were used here.

C. Results

Figure 7 shows the two glimpse measures for each of the conditions used in the experiments of the previous section. For the stationary noise conditions corresponding to experiment I, significantly more glimpses (as measured by both area and proportion) were produced by the noise-induced speech than for speech produced in quiet ($p < 0.01$). Stationary noise maskers produced more glimpses than the competing speech [$F(1,7)=14.501$, $p < 0.001$ for area; $F(1,7)=10.513$, $p < 0.01$ for proportion] while the effect of an increase in noise level was significant only for the glimpse area measure [$F(1,7)=6.438$, $p < 0.05$]. Regarding the two competing talker conditions of experiment II, both show significantly more glimpses than speech produced in quiet when measured in terms of glimpse area [$t(7)=3.398$, $p < 0.05$ for the less intense condition; $t(7)=3.780$, $p < 0.01$ for the more intense condition] while there is a small increase in glimpse

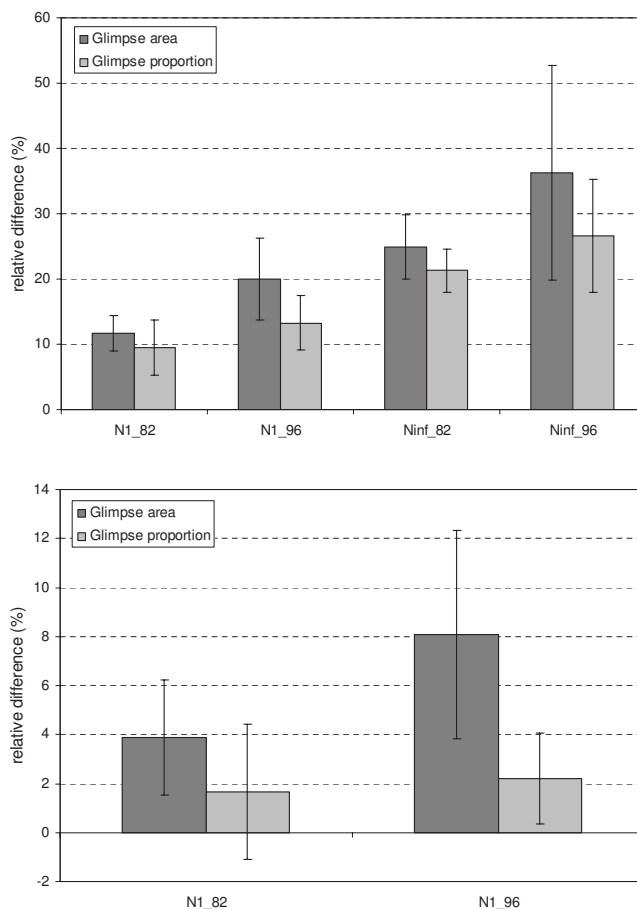


FIG. 7. Glimpse area and proportion for the listening conditions of experiments I (top) and II (bottom) expressed as percentage increase in area or proportion over speech produced in quiet.

proportion for the more intense condition [$t(7)=2.384$, $p < 0.05$]. Finally, as was the case for intelligibility, no significant effect was found for experiment III.

Overall, the results are strikingly similar to those for intelligibility, as illustrated by Fig. 8 which plots relative intelligibility gains for listeners against relative increases in the two glimpse measures. Both measures are highly correlated with listener intelligibility gains, suggesting that noise-

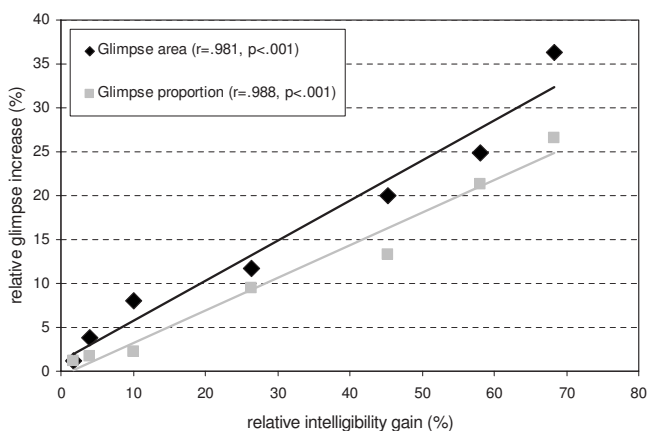


FIG. 8. Relation between increases in the two glimpse measures and increase in intelligibility for experiments I–III together with least-squares fits.

induced speech is more intelligible than speech produced in quiet because the articulatory manipulations lead to a release from EM.

Of the two glimpse measures, significantly larger increases in glimpse area over glimpse proportion are found [$F(1,7)=4.102$, $p<0.05$; $F(1,7)=7.397$, $p<0.05$] for the conditions of both experiments I and II. This is presumably due to the tendency of noise-induced sentences to increase in duration. No significant correlation was found between utterance-wise measures of duration and glimpse proportion in any of the noise-induced conditions, suggesting that speakers use both a slower speaking rate, to increase the overall number of glimpses, and other (mainly spectral) modifications in order to increase the proportion of glimpses available for the hearer. However, in explaining listener performance, there is no clear basis to prefer glimpse area over glimpse proportion.

VI. GENERAL DISCUSSION

A. Noise-induced speech and energetic masking

A number of reliable and consistent acoustic modifications occur when speech is produced in the presence of noise. The main effects—increases in F0, energy, and spectral CoG—confirm those found in previous work using multitalker babble and stationary noise. The study extends the scope of noise-induced production effects to single-talker interfering speech, which was also found to be capable of producing significant acoustic changes compared to speech produced in a quiet background. Increases in F0, energy, and CoG grew as the number of talkers in the background increased, asymptoting at around 8–16 talkers. These results demonstrate that the extent of acoustic modifications is largely correlated with both the intensity of the background signal and number of background talkers. This suggests that noise-induced speech production changes are dependent on the overall EM capacity of the background signal since EM is a function of both overall noise level and number of background talkers: a competing talker is a far less effective energetic masker than a broadband noise when both are presented at the same SNR (Festen and Plomp, 1990), and EM increases with the number of background talkers (Bronkhorst and Plomp, 1992; Simpson and Cooke, 2005).

Experiment I demonstrated that noise-induced speech was more intelligible when presented in stationary noise than speech produced in quiet, extending previous findings for stationary and multitalker babble backgrounds. Interestingly, those backgrounds which resulted in the largest acoustic modifications also produced the biggest increases in intelligibility, suggesting that speakers modify their productions in response to the adversity of the background. Indeed, the result of production modifications is to increase the number and proportion of opportunities to glimpse the target speech in noise, and the increase in such opportunities is very closely correlated with listener keyword identification performance. Thus, the potential for EM leads to articulatory modifications whose acoustic consequence is to cause a release from masking, and the more masking potential that exists,

the greater the eventual release. These findings support Lindblom's suggestion that speakers compensate for environmental conditions (Lindblom, 1990).

B. Basis for the increased intelligibility of noise-induced speech

It is not clear how the acoustic consequences of changes in speech production lead to increased intelligibility. While it is evident that the overall increase in intensity of noise-induced speech produces a release from EM, this cannot account for the intelligibility gains observed here since all utterances were normalized to have the same SNR when presented alongside maskers. However, speakers can employ a number of other strategies to improve the SNR at the ear of the listener. For instance, a decrease in speaking rate provides more opportunities to glimpse acoustic information useful for phonetic distinctions. The largest increase in utterances duration of around 7% in the most adverse backgrounds might have contributed to the overall improvement in intelligibility, but it is unlikely to be responsible for the entire increase since the results of the glimpsing analysis showed that the *proportion* of the spectrum lying above the masker also increased for the noise-induced conditions.

Many of the acoustic consequences of noise-induced speech are compatible with an overall shift in the energy balance from lower to higher frequencies. For the vowels, increases in fundamental frequency, spectral CoG, and energy for the second and third formants are reflected in a flattening of spectral tilt. One consequence of this shift to higher frequencies is a certain degree of masking release in the presence of the maskers employed in this study, whose mean spectrum was speech shaped. However, vowel formant frequencies became more “central,” with increases in F1 and decreases in F3. Of course, there are articulatory limits to the range of speech production modifications possible, and some of the acoustic changes observed may be epiphenomena associated with other manipulations such as increased effort and vocal stress.

The issue of whether speakers actively attempt to place spectral information in locations where it is less likely to be masked merits further study. Formant frequency changes are also found when talkers are asked to speak clearly (Chen, 1980; Picheny *et al.*, 1986; Krause and Braida, 2004; Smiljanic and Bradlow, 2005). These studies categorized vowels as tense or lax, corresponding to /i:/, u:/ and /ɪ, e/ here. No consistent trends were found for the first three formant frequencies of tense vowels, although Chen (1980) reported that tense vowels clustered more tightly in vowel formant space in clear than in conversational speech. Picheny *et al.* (1986) and Krause and Braida (2004) reported increases in F1 and F2 of the lax vowel /ɪ/ of up to 50 and 200 Hz, respectively.

C. Speech changes produced by a competing talker

One of the motivations for the current study was to determine how the presence of a competing talker affects speech production. One possibility is that competing speech material might disrupt the speech production process of the

talker, resulting in false starts, hesitations, and other dysfluencies. The speech material used in this study was deliberately chosen to be similar to that introduced in the background in order to provoke such effects. Some disrupting influence of the competing talker background was found: the number and duration of short pauses increased with intensity while no similar effects were seen for the stationary noise backgrounds. Furthermore, the number of false starts was larger in the intense single-talker background than in quiet [$t(7)=2.646$, $p<0.05$]. However, these effects were small and the overall number of short pauses was not significantly greater than in a quiet background.

A second potential influence of competing speech is on the talker-listener communication process: the talker might anticipate the IM effect of two similar utterances at the ear of the listener and employ strategies to reduce the extent of IM. Experiment II demonstrated that utterances produced in the presence of an intense competing talker were more intelligible than utterances produced in quiet conditions when presented in speech backgrounds. For speech produced with a less intense talker, there was no significant gain over quiet. These findings suggest that it is primarily EM rather than IM that leads to increased intelligibility since if the latter were at work, some effect in the less intense background would be expected since the production and background levels are closer and lead to more IM for the listeners (Brungart, 2001). Furthermore, no evidence was found of speaking strategies which exploited the temporal fluctuations of specific competing utterances: there was no difference in the intelligibility of speech in the presence of the material which induced it when compared to speech in the presence of other speech material (experiment III). Talkers may be unable to attend to and track competing speech material sufficiently rapidly to modify their own productions in response.

D. Task dependence

While few effects of a competing talker above and beyond EM were found here, it is possible that other tasks might elicit more extensive speech production changes. The task employed in the current study was devoid of communicative intent, and it was possible for speakers to read the prompts on the screen with little regard for intelligibility. Summers *et al.* (1988) found that with no communicative element there was little incentive for speakers to consciously change their speech even with masking noise present in the headphones. Lane and Tranel (1971) indicated that the speaker does not change his voice level to communicate better with himself, but rather with others. Junqua *et al.* (1999) also concluded that the communication factor has a strong influence. Further studies using two-way interactive task with single-talker maskers need to be conducted before ruling out the possibility of both positive effects of active strategies which are sensitive to the local masking conditions and negative effects of attentional deployment to processing an informative background source while speaking.

It is known that the greatest IM effects are found when the target signal and masker are similar (Brungart, 2001; Neff, 1995; Oh and Lutfi, 2000; Kidd *et al.*, 2002). Indeed,

although the masking utterances were similar in form to those produced by the talkers, start times were not synchronized, so the chances of similar words overlapping was reduced. It is possible that tasks designed to produce large amounts of IM would give rise to more significant changes in speech production than those observed in the current study.

VII. CONCLUSIONS

Speakers modified their productions in N -talker noise backgrounds across a wide range of values for N . This they achieved not only by increases in output level but by changes to the fundamental frequency and formant energies which result in an overall increase in spectral CoG. The scale of acoustic modifications increased both with N and the level of the background noise, conditions which also result in increases in the EM effect of the noise. Noise-induced speech was more intelligible when presented in stationary noise than speech produced in quiet, and the intelligibility gain increased with N and noise level. These findings, coupled with a computer model of EM, suggest that speakers attempt to compensate for the EM effect of the noise on their own speech. In contrast, no IM effects of a competing talker were found, perhaps because the task lacked a communicative element.

APPENDIX: COMPENSATION FOR OWN-VOICE ATTENUATION

To determine whether own-voice attenuation caused by closed headphones was a factor in the current study, a compensation method was introduced. First, the spectral difference of a white noise signal with and without Sennheiser HD 250 Linear II headphones was measured using a B & K type 4100 head and torso simulator equipped with B & K type 4190 $\frac{1}{2}$ in. microphones. An order-32 IIR filter was designed to have a transfer function which was the inverse of the attenuation characteristic produced by the headphones. This filter was implemented on a TDT RP 2.1 processor and compensated for the headphone attenuation in real time.

In order to discover whether the original and the compensated recording method produced similar effects on speech production, a small corpus was collected using the two methods and analyzed at utterance and phoneme level. Eight native speakers of British English (four males and four females) drawn from staff and students in the Department of Computer Science at the University of Sheffield participated in the corpus collection. Eight recording conditions were employed which included quiet, competing talker, eight-talker babble, and speech-shaped noise. Talkers produced the same set of 25 sentences in each of the eight conditions. Maskers for noise conditions were produced as described in Sec. II B and presented at 89 dB SPL. Condition order was randomized for each talker. For the collected utterances, leading and trailing silent intervals identified via the alignment process described in Sec. II E were removed.

Four acoustic properties were estimated for each utterance in each of the eight conditions. Sentence duration, rms energy, mean fundamental frequency (F_0), and spectral CoG

were computed as described in Sec. III A. A two-way repeated-measures ANOVA (two recording methods \times three noise conditions) was computed for each acoustic parameter. *Post hoc* analysis showed that for all parameters and noise conditions, there was no significant effect of recording method [$F(1,7)=0.893$, $p=0.376$ for duration; $F(1,7)=1.240$, $p=0.278$ for energy; $F(1,7)=0.923$, $p=0.369$ for F0; $F(1,7)=1.055$, $p=0.339$ for CoG]. For the quiet and competing talker conditions, short pauses within each utterance were manually identified and their number and duration computed. Again, the difference in recording setups led to no statistically significant differences [$F(1,7)=0.007$, $p=0.936$ for the number of short pauses; $F(1,7)=0.056$, $p=0.820$ for duration of short pauses].

¹However, speech produced in the presence of other speech material has been studied in the limited sense of altered auditory feedback (Lee, 1950; Natke and Kalverman, 2001; Stuart *et al.*, 2002; Xu *et al.*, 2004).

²Processing delays in the TDT System 3 processor mean that the noise output was slightly delayed (maximum 6 ms) with respect to speech input.

³Mean energy was computed via “get intensity decibels.” F0 estimates were provided at 10 ms intervals using an autocorrelation-based method (Boersma, 1993) implemented in the PRAAT program. Mean F0 was obtained by averaging all the valid F0 estimates and expressed in semitones. Spectral CoG was computed on the spectrum of an entire utterance via the PRAAT command “get CoG” and expressed in hertz, using a linear frequency axis and power magnitude spectrum.

Arlinger, S. D. (1986). “Sound attenuation of TDH-39 earphones in a diffuse field of narrow-band noise,” *J. Acoust. Soc. Am.* **79**, 189–191.

Barker, J., and Cooke, M. P. (2007). “Modelling speaker intelligibility in noise,” *Speech Commun.* **49**, 402–417.

Boersma, P. (1993). “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” *Proc. Inst. Phonetic Sci.* **17**, 97–110.

Boersma, P., and Weenink, D. (2005). “Praat: doing phonetics by computer (version 4.3.14) (computer program),” (Last viewed May, 2005) from <http://www.praat.org>.

Bond, Z., Moore, T., and Gable, B. (1989). “Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask,” *J. Acoust. Soc. Am.* **85**, 907–912.

Bořil, H., Bořil, T., and Pollák, P. (2006). “Methodology of Lombard speech database acquisition: Experiences with CLSD,” LREC 2006 fifth Conference on Language Resources and Evaluation, pp. 1644–1647.

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.

Carhart, R., Johnson, C., and Goodman, J. (1975). “Perceptual masking of spondees by combinations of talkers,” *J. Acoust. Soc. Am.* **58**, S35.

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

Charlip, W. S., and Burk, K. W. (1969). “Effects of noise on selected speech parameters,” *J. Commun. Dis.* **2**, 212–219.

Chen, F. R. (1980). “Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level,” MS thesis, Massachusetts Institute of Technology, Cambridge.

Chi, S. M., and Oh, Y. H. (1996). “Lombard effect compensation and noise suppression for noisy Lombard speech recognition,” International Conference on Spoken Language Processing, pp. 2013–2016.

Cooke, M. P. (2006). “A glimpsing model of speech perception in noise,” *J. Acoust. Soc. Am.* **119**, 1562–1573.

Cooke, M. P., Barker, J., Cunningham, S., and Shao, X. (2006). “An audio-visual corpus for speech perception and automatic speech recognition,” *J. Acoust. Soc. Am.* **120**, 2421–2424.

Cooke, M. P., Garcia Lecumberri, M. L., and Barker, J. P. (2008). “The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception,” *J. Acoust. Soc. Am.*

123, 414–427.

Dreher, J. J., and O’Neill, J. (1957). “Effects of ambient noise on speaker intelligibility for words and phrases,” *J. Acoust. Soc. Am.* **29**, 1320–1323.

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.

Garnier, M., Bailly, L., Dohen, M., Welby, P., and Loevenbruck, H. (2006). “An acoustic and articulatory study of Lombard speech: Global effects on the utterance,” International Conference on Spoken Language Processing, pp. 2246–2249.

Gramming, P., Sundberg, J., Ternstöm, S., Leanderson, R., and Perkins, W. (1988). “Relationship between changes in voice pitch and loudness,” *J. Voice* **2**, 118–126.

Hanley, T. D., and Steer, M. D. (1949). “Effect of level of distracting noise upon speaking rate, duration, and intensity,” *J. Speech Hear. Disord.* **14**, 363–368.

Hansen, J. H. L. (1994). “Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect,” *IEEE Trans. Speech Audio Process.* **2**, 598–614.

Hansen, J. H. L. (1996). “Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition,” *Speech Commun. special issue on speech under stress*, **20**, 151–170.

Hansen, J. H. L., and Bria, O. N. (1990). “Lombard effect compensation for robust automatic speech recognition in noise,” International Conference on Spoken Language Processing, pp. 1125–1128.

Howes, D. (1957). “On the relation between the intelligibility and frequency of occurrence of English words,” *J. Acoust. Soc. Am.* **29**, 296–305.

Junqua, J. C. (1993). “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *J. Acoust. Soc. Am.* **93**, 510–524.

Junqua, J. C. (1994). “A duration study of speech vowels produced in noise,” International Conference on Spoken Language Processing, pp. 419–422.

Junqua, J. C. (1996). “The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex,” *Speech Commun.* **20**, 13–22.

Junqua, J. C., Fincke, S., and Field, K. (1998). “Influence of the speaking style and the noise spectral tilt on the Lombard reflex and automatic speech recognition,” International Conference on Spoken Language Processing, pp. 467–470.

Junqua, J. C., Fincke, S., and Field, K. (1999). “The Lombard effect: A reflex to better communicate with others in noise,” *Acoustic, Speech, and Signal Processing, 1999 (ICASSP’99) Proceedings, Vol. 4*, pp. 2083–2086.

Kidd, G., Mason, C. R., and Arbogast, T. L. (2002). “Similarity, uncertainty and masking in the identification of nonspeech auditory patterns,” *J. Acoust. Soc. Am.* **111**, 1367–1376.

Krause, J. C., and Braid, L. D. (2004). “Acoustic properties of naturally produced clear speech at normal speaking rates,” *J. Acoust. Soc. Am.* **115**, 362–378.

Lane, H. L., and Tranel, B. (1971). “The Lombard sign and the role of hearing in speech,” *J. Speech Hear. Res.* **14**, 677–709.

Lee, B. S. (1950). “Effects of delayed speech feedback,” *J. Acoust. Soc. Am.* **22**, 824–826.

Letowski, T., Frank, T., and Caravella, J. (1993). “Acoustical properties of speech produced in noise presented through supra-aural earphones,” *Ear Hear.* **14**, 332–338.

Lindblom, B. (1990). “Explaining phonetic variation: A sketch of the H&H theory,” in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer Academic, The Netherlands), pp. 403–439.

Lindblom, B., and Sundberg, J. (1971). “Acoustical consequences of lip, tongue, jaw, and larynx movement,” *J. Acoust. Soc. Am.* **50**, 1166–1179.

Lombard, E. (1911). “Le Signe de l’Elevation de la Voix (The sign of the rise in the voice),” *Ann. Maladies Oreille, Larynx, Nez, Pharynx (Annals of diseases of the ear, larynx, nose and pharynx)*, **37**, 101–119.

Natke, U., and Kalveram, K. T. (2001). “Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables,” *J. Speech Lang. Hear. Res.* **44**, 1–8.

Neff, D. L. (1995). “Signal properties that reduce masking by simultaneous random-frequency maskers,” *J. Acoust. Soc. Am.* **98**, 1909–1920.

- Oh, E. L., and Lutfi, R. A. (2000). "Effect of masker harmonicity on informational masking," *J. Acoust. Soc. Am.* **108**, 706–709.
- Patel, R., and Schell, K. W. (2008). "The influence of linguistic content on the Lombard effect," *J. Speech Lang. Hear. Res.* **51**, 209–220.
- Picheny, M. A., Durlach, N. I., and Braida, L. D. (1986). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *J. Speech Hear. Res.* **29**, 434–446.
- Pickett, J. M. (1956). "Effects of vocal force on the intelligibility of speech sounds," *J. Acoust. Soc. Am.* **28**, 902–905.
- Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C., and Yuchtman, M. (1985). "Some acoustic-phonetic correlates of speech produced in noise," *International Conference on Acoustics Speech and Signal Processing*, pp. 1581–1584.
- Pittman, A. L., and Wiley, T. L. (2001). "Recognition of speech produced in noise," *J. Speech Lang. Hear. Res.* **44**, 487–496.
- Simpson, S. A., and Cooke, M. P. (2005). "Consonant identification in *N*-talker babble is a nonmonotonic function of *N*," *J. Acoust. Soc. Am.* **118**, 2775–2778.
- Smiljanic, R., and Bradlow, A. R. (2005). "Production and perception of clear speech in Croatian and English," *J. Acoust. Soc. Am.* **118**, 1677–1688.
- Stanton, B., Jamieson, L., and Allen, G. (1988). "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions," *International Conference on Acoustics Speech and Signal Processing*, pp. 331–334.
- Steeneken, H. J. M., and Hansen, J. H. L. (1999). "Speech under stress conditions: Overview of the effect on speech production and on system performance," *International Conference on Acoustics Speech and Signal Processing*, pp. 2079–2082.
- Stuart, A., Kalinowski, J., Rastatter, M. P., and Lynch, K. (2002). "Effect of delayed auditory feedback on normal speakers at two speech rates," *J. Acoust. Soc. Am.* **111**, 2237–2241.
- Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (1988). "Effects of noise on speech production: Acoustic and perceptual analysis," *J. Acoust. Soc. Am.* **84**, 917–928.
- Tartter, V. C., Gomes, H., and Litwin, E. (1993). "Some acoustic effects of listening to noise on speech production," *J. Acoust. Soc. Am.* **94**, 2437–2440.
- Varadarajan, V. S., and Hansen, J. H. L. (2006). "Analysis of Lombard effect under different types and levels of noise with application to in-set speaker ID system," *International Conference on Spoken Language Processing*, pp. 937–940.
- Webster, J. C., and Klumpp, R. G. (1962). "Effects of ambient noise and nearby talkers on a face-to-face communication task," *J. Acoust. Soc. Am.* **34**, 936–941.
- Womack, B., and Hansen, J. (1996). "Classification of speech under stress using target driven features," *Speech Commun.*, special issue on speech under stress, **20**, 131–150.
- Xu, Y., Larson, C. R., Bauer, J. J., and Hain, T. C. (2004). "Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences," *J. Acoust. Soc. Am.* **116**, 1168–1178.
- Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (1999). *The HTK Book 2.2* (Entropy, Cambridge).