Speaking in the presence of a competing talker

Youyi Lu[1], *Martin Cooke*[2,3]

[1]Department of Computer Science, University of Sheffield, UK
[2]Ikerbasque (Basque Science Foundation)
[3]Language and Speech Laboratory, Faculty of Letters, Universidad del País Vasco, Spain
y.lu@dcs.shef.ac.uk, m.cooke@ikerbasque.org

Abstract

How do speakers cope with a competing talker? This study investigated the possibility that speakers are able to retime their contributions to take advantages of temporal fluctuations in the background, reducing any adverse effects for an interlocutor. Speech was produced in quiet, competing talker, modulated noise and stationary backgrounds, with and without a communicative task. An analysis of the timing of contributions relative to the background indicated a significantly reduced chance of overlapping for the modulated noise backgrounds relative to quiet, with competing speech resulting in the least overlap. Strong evidence for an active overlap avoidance strategy is presented.

Index Terms: competing talker, speech production, Lombard effect, glimpsing, temporal overlap

1. Introduction

A common experience in today's mobile-phone-dominated world is finding oneself having a conversation in the presence of other competing talkers. Surprisingly, while the effect of broadband noise and multi-talker babble on speech production has been investigated [1, 2, 3, 4], there have been very few studies when the background contains a clearly audible speech signal. Yet such backgrounds might be expected to lead to changes different from the classic *Lombard* effects (e.g. increases in speech output power, F0 and spectral centre of gravity) observed as a response to broadband noise.

One interpretation of Lombard speech suggests that talkers attempt to compensate for the masking effect of the noise at the ear of the listener. A recent study [5] demonstrated that as the level or spectral and temporal density of the noise increases, talkers appear to create more opportunities for listeners to glimpse the target speech [6]. This could be a response to the *energetic* masking (EM) effect of the noise i.e. masking caused by the overlap of energy from more than one sources in the auditory periphery. A competing talker also produces linguistic or *informational* masking (IM) due to the difficulty in segregating similar talkers, amongst other causes [7]. It is possible that talkers not only attempt to overcome EM for the listener, but also engage in speech production strategies that minimise the degree of IM.

One of the few studies of speech in the presence of other talkers was carried out by [8], in which talker-listener pairs were seated face to face and communicated word lists in quiet and in the presence of 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, and the foreground pair made more communication errors when talking at the same time as the competing pair.

A single talker background was also used in [5], who found that when equalised for level, a competing talker led to smaller Lombard effects than those produced by stationary noise. Little evidence of production changes related to IM was found. However, that study involved the reading of sentence prompts, and it is likely that a communicative task would produce stronger effects [3].

The purpose of the current study was to examine the temporal structure of speech produced in the presence of a fluctuating masker. In particular, we were interested in whether talkers could avoid temporal overlap with the masker to ameliorate both EM and IM for listeners. Section 2 outlines the task, corpus and noise backgrounds, while sections 3 and 4 respectively describe traditional Lombard and temporal effects of stationary and fluctuating maskers on speech production.

2. Speech corpus

2.1. Tasks

Two tasks involving the solution of Sudoku puzzles were employed. In the non-communicative task, individual speakers were asked to speak aloud while tackling the puzzles, while in the communicative task pairs of speakers solved these puzzles cooperatively. Sudoku puzzles were chosen because they provoke the use of spoken digits which serve as a robust basis for across-condition comparisons. The background was quiet (Q) or contained one of three types of maskers: competing speech (CS), speech-modulated noise (SMN) or speech-shaped noise (SSN). Speech-modulated noise provides the same opportunities for exploitation of temporal gaps as natural speech since it has an identical temporal envelope, but it has the spectrum of stationary speech-shaped noise and is not intelligible.

Speech was collected from eight native speakers of British English (4 males and 4 females), grouped into 4 pairs. Each speaker attended three recording sessions. In the first, without noise exposure, one speaker in each pair did a 10 minute recording while solving puzzles alone. Then, the speaker cooperated with his/her partner for 10 minutes, followed by another 10 minute recording when the partner was solving alone. From this material, speech from 2 males and 2 females was selected to be used as the basis for competing speech maskers in subsequent sessions. Ten minutes of speech from each of the 4 talkers was manually transcribed to identify speech/nonspeech segments and silent pauses. Sound types such as *uh*, *um*, *ooh*, paper-rustle, breathing, laughing, coughing, and unintelligible utterances were labelled as *nonspeech*. Silent pauses longer

than 100 ms were also identified. Each nonspeech segment was replaced with a silence of the same duration. The resulting four signals were used as the competing speech maskers. For each competing speech masker, a speech-shaped noise signal was generated by filtering white noise with a filter whose spectrum equalled the long-term spectrum of the speech segments of the competing masker, and the corresponding speech-modulated noise was formed by modulating the generated speech-shaped noise with the envelope of the competing speech masker.

Speakers participated in two further sessions on different days in which they solved puzzles alone (session 2) and with their partners (session 3) in the three noise conditions. Recordings lasted 10 minutes for each noise condition. In the third session, for each pair of speakers, the masker used in the competing speech condition contained a talker of the same gender, since this is known to provoke more IM [7].

2.2. Recording

Corpus collection sessions took place in an IAC single-walled acoustically-isolated booth. When working together, each pair of talkers sat at two sides of a table which had a barrier screen in the middle, providing some acoustic isolation to reduce crosstalk and to prevent eye contact forcing talkers to rely only on acoustic cues. Two Bruel & Kjaer (B & K) type 4190 $\frac{1}{2}$ inch microphones each coupled with a preamplifier (B & K type 2669) were fixed on the screen and directed towards each talker. When seated, the distance between the talker and the nearest microphone was set at approximately 20 cm.

Each talker's signal was further processed by a conditioning amplifier (B & K Nexus model 2690) prior to digitisation at 25 kHz with a Tucker-Davis Technologies (TDT) System 3 RP2.1. Simultaneously, maskers were presented diotically over Sennheiser HD 250 Linear II headphones using the same TDT system at 82 dB SPL, a level selected within the range known to provide sufficient energetic masking (e.g. [1] used 80, 90 and 100 dB; [2] used 85 dB) but still relatively low in order to elicit informational masking effects. It is known that listeners can exploit level differences between talkers [7], so an intense masker might conceivably reduce IM. Talkers wore headphones throughout, including for the quiet condition.

2.3. Transcription

Recordings were manually transcribed at two different levels: (1) speech/nonspeech segments and silent pauses (>100 ms); and (2) individual digits "one" to "nine". There were on average 12.3 instances (s.d.=4.2) of each digit available in each condition per talker.

3. Lombard effects

The primary purpose of the current study was to examine the temporal effects of speaking in the presence of modulated noise backgrounds. However, we also wished to confirm the existence and size of traditional Lombard effects. Fig. 1 illustrates speech energy, F0, and spectral tilt as a function of task and background. For both tasks, noise backgrounds led to an increase in energy and mean F0 and a decrease in spectral tilt, with the greatest effect for SSN [$F(3,21)=17.98$, $p<0.001$, $\eta^2=0.72$ for energy; $F(3,21)=8.98$, $p<0.01$, $\eta^2=0.56$ for F0; $F(3,21)=7.70$, $p<0.01$, $\eta^2=0.52$ for spectral tilt]. There was also a clear task effect which led to more extreme changes in energy and spectral tilt [$F(1,7)=26.08$, $p<0.01$, $\eta^2=0.79$ for energy; $F(1,7)=28.57$, $p<0.01$, $\eta^2=0.80$ for spectral tilt] and a smaller effect for F0.

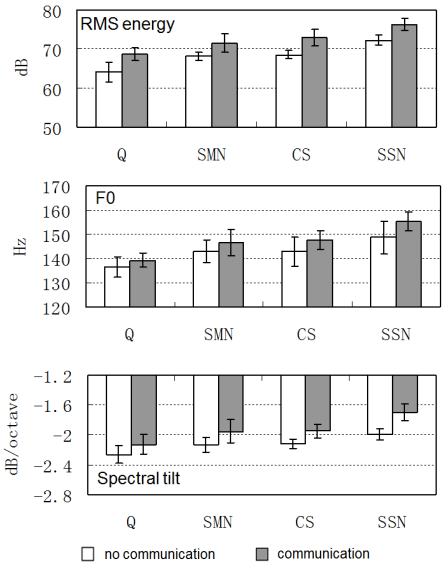

Figure 1: *RMS energy, F0 and spectral tilt as a function of task and background. Values shown are means over talkers and error bars indicate 95% confidence intervals, here and elsewhere.*

A competing talker and speech-modulated noise resulted in similar Lombard effects whether the communication factor was present or not. Given that these two maskers yield approximately equal EM, this result is compatible with the hypothesis that the size of speech production changes scales with the EM potential of the background noise [5]. In response to noise, an increase in speech level can benefit speech intelligibility due to an increase in signal-to-noise ratio, as well as the flattening of spectral slope which enables more of the speech to escape masking, at least for the maskers used here which had a low-frequency bias. On the other hand, increases in F0 might be correlated with a change in speech level, and have been found to contribute little to speech intelligibility [9]. These findings extend the results of [5] using read sentences to a task involving communication.

The increase of speech level and the shift of spectral energy towards higher frequencies produced by the communication factor in the presence of noise were found in [3]. However, unlike [3] there was no *additional* effect of communication on the size of the noise-induced speech production changes in speech level, F0 and spectral tilt. It is unclear whether this discrepancy is due to task differences or other factors.

4. Temporal effects

4.1. Foreground-background overlap

Here, the issue of whether talkers could avoid overlapping in time with a noise background was studied by measuring the length of temporal overlap between speech activity in the foreground talker and speech or speechlike (in the case of SMN) activity in the background masker. The overlap values were computed relative to the length of speech from the foreground, expressed as overlap percent, in order to normalise for differences in the amount of speech produced across conditions.

For each talker, the overlap was computed between the *foreground* speech segments produced in the backgrounds with temporal fluctuations (CS and SMN) and the *background* in which

the speech was collected (fig. 2). As a reference, for each talker, the overlap between speech segments produced in quiet and the background used in the fluctuating masker case was also computed. If talkers were attempting to make use of the gaps in the fluctuating background, one would expect to see a smaller degree of overlap relative to the quiet case.

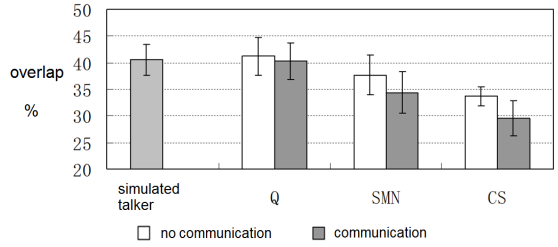

Figure 2: *Overlap percent as a function of task and background. The leftmost bar shows the degree of overlap for a simulated talker (see section 4.4).*

Compared to quiet, both tasks produced a significant reduction in overlap in fluctuating noise conditions, with a greater reduction for competing speech [F(2,14)=44.82, p<0.001, η^2=0.87]. In addition, compared to the task with no communication, the communicative task led to a significantly smaller overlap percentage in the backgrounds of SMN [p<0.05] and CS [p<0.01] but not in quiet [p=0.25].

There are a number of ways in which speakers could reduce foreground-background overlap in CS and SMN relative to quiet. It is possible that talking more rapidly or changing pause length distribution might result in overlap reduction *without any active attempt* to time contributions relative to the background. Subsequent analyses addressed these issues.

4.2. Speaking rate

The mean speaking rate in each condition and for each talker was estimated using the digits extracted during corpus transcription. To accommodate the different numbers of digit exemplars in each condition, a certain number n_i of each of the digits $i = 1..9$ (different for each digit but fixed across conditions) was chosen and speaking rate $rate_c$ for condition c was computed according to:

$$rate_c = \frac{\sum_{i=1}^{9} n_i}{\sum_{i=1}^{9} \sum_{k=1}^{n_i} d_{cik}} \quad (1)$$

where d_{cik} is the duration of the kth exemplar of digit i in condition c. Fig. 3 shows a clear increase in speaking rate for the communicative task [F(1,7)=28.44, p<0.01, η^2=0.80] but no effect of noise background. Thus, speaking rate changes cannot account for the overlap reduction as a function of noise background. Further, while the difference in speaking rate across tasks might at first sight be considered as a contributory factor given the task differences in fig. 2, this is unlikely since in the quiet condition there was no task effect on overlap yet the task produced a significantly faster speaking rate.

4.3. Mean pause duration

Another factor which could lead to reduced overlap is a change in pause structure as a function of the background or task. Mean pause durations (fig. 4) do indeed show both task and background effects. The communicative task resulted in longer

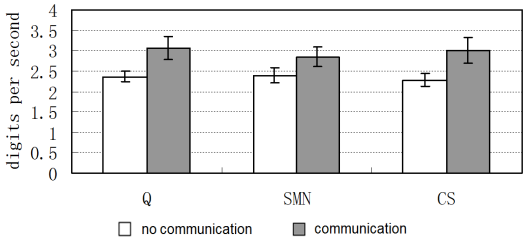

Figure 3: *Speaking rate as a function of task and background.*

pauses overall [F(1,7)=9.70, p<0.05, η^2=0.58], although not in quiet. Both tasks showed longer pauses in the modulated noise conditions. For the communicative task, this trend was statistically significant [F(1.98,13.88)=9.04, p<0.01, η^2=0.56]. Comparison of figs 2 and 4 reveals a common pattern. Longer pause durations correlate strongly with decreasing amounts of overlap [ρ = -0.90, p<0.05]. This finding is consistent with the idea that speakers wait until an appropriate point to make their contributions in the face of a modulated background. However, it is also possible that the mere presence of noise results in longer pauses. The rightmost bars of fig. 4 suggest otherwise. The mean pause duration for stationary noise is barely different from quiet [p>0.05]. SSN produces the largest Lombard effects (fig. 1) but has little effect on pause duration.

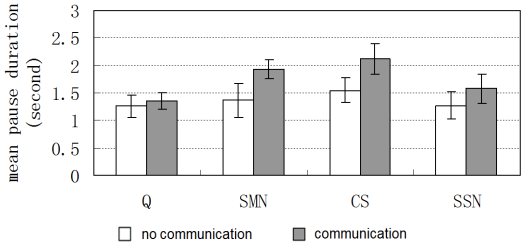

Figure 4: *Mean pause durations.*

4.4. Simulated talkers

There remains the possibility that the pause distribution varies as a function of the background (e.g. speakers matching their rhythm to that of a competing talker) without necessarily requiring *active* timing of contributions to avoid overlap. To test this idea, a simulated talker having the same distribution of pause and contribution lengths as the real talkers was constructed.

Example distributions of pause and contribution lengths for a single talker in quiet and competing speaker backgrounds are shown in fig. 5. To accommodate the long one-sided tail, gamma distributions with density given by

$$f(x;\alpha;\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \quad (2)$$

parameterised by α ("shape") and β^{-1} ("rate") were fitted to each pause and contribution distribution. A talker's pause structure in each condition was then simulated by alternately sampling from the gamma distributions for pauses and contributions to produce a sequence of the same length as the real speaker data. Fifty simulation sequences were produced for each condition.

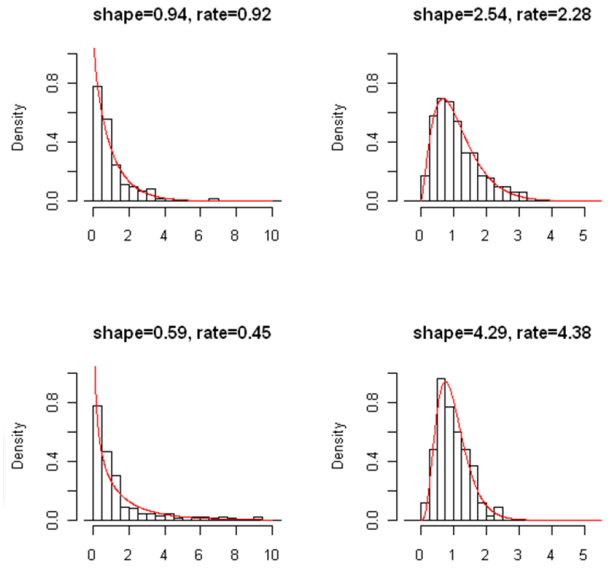

Figure 5: *Pause length (left) and contribution length (right) densities for a single talker in quiet (top) and a competing talker background (bottom). Horizontal axis is duration in seconds. Gamma fits are also plotted with shape and rate values shown.*

The overlap rates for these *simulated* talkers were statistically-identical across tasks and noise backgrounds. The degree of overlap for the simulated talkers is plotted in fig. 2 and matches very closely the real talker data in the quiet condition. An additional simulation was performed by randomising the order of consecutive pause-contribution pairs from the original data. Again, overlap scores (40%) similar to those in quiet were obtained. These simulations demonstrate that random sampling from the different pause and contribution duration distributions cannot account for the differences in overlap rate across the tasks and backgrounds.

5. Discussion

The key finding of the current study was that speakers attempt to avoid overlapping with fluctuating noise backgrounds. The reduction in overlap could not be accounted for by "passive" factors such as speaking rate changes or simulated talkers with identical pause distributions as natural talkers. The reduction was greater for competing speech than for speech modulated noise, and greater for the communicative task. This, as far as we are aware, is the first report of active speaking behaviour in response to noise.

Avoidance of temporal overlapping leads to a release from energetic masking, aiding segregation of foreground and background speech for the interlocutor. The additional overlap reduction produced by the competing talker background relative to the speech-modulated noise may also result in reduced information masking due to improved foreground-background segregation [10]. The perceptual mechanisms which drive the reduction in overlap are unclear. One possibility is that intelligibility of the competing speech masker relative to the speech-modulated noise allows a better prediction of upcoming pauses. This strategy is supported by the data of figs 4 and 5: for the competing talker background, there is evidence that the increased mean pause duration is largely due to a greater number of long pauses, perhaps due to speakers' monitoring the background for a suitable place to interject.

While the current results showed the possible presence of a temporal-domain strategy to yield a release from energetic and informational masking, there are other mechanisms open to speakers. For example, differences in speech level or F0 between foreground and background are known to reduce informational masking [7]. In the present study, observed changes in speech level and F0 in the competing speech condition appeared to be governed primarily by energetic masking factors since speech-modulated noise induced very similar level and F0 changes. It may be that temporal domain speech manipulation is an efficient form of talker behaviour compared to manipulations of vocal level and F0. Increasing speech level is energy consuming and the extent to which talkers can manipulate F0 is constrained by articulatory constraints [11].

Overall, these findings suggest that talkers adopt a "listening-while-speaking" strategy which helps to increase the probability of message reception at the ears of the interlocutor. Most of the benefit arises from a reduction in energetic masking, by both spectral and temporal reallocation of speech energy to frequency regions and time intervals where it is least likely to be masked. Further studies will reveal whether speakers adopt other foreground-background "misalignment" strategies to help in communication.

Acknowledgements. This work was supported by the Marie Curie Research Training Network "Sound to Sense".

6. References

[1] W. Summers, D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production: Acoustic and perceptual analysis," *J. Acoust. Soc. Am.*, vol. 84, pp. 917–928, 1988.

[2] J. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, pp. 510–524, 1993.

[3] M. Garnier, "Communiquer en environnement bruyant: de l'adaptation jusqu'au forçage vocal [communication in noisy environments: from adaptation to vocal straining]," These de Doctorat de l'Universite Paris 6, 2007.

[4] H. Bořil, "Robust speech recognition: analysis and equalization of lombard effect in czech corpora," Doctoral thesis, Czech Technical University, Prague, 2008.

[5] Y. Lu and M. P. Cooke, "Speech production modifications produced by competing talkers, babble and stationary noise," *J. Acoust. Soc. Am.*, vol. 124, pp. 3261–3275, 2008.

[6] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[7] D. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 109, pp. 1101–1109, 2001.

[8] J. Webster and R. Klumpp, "Effects of ambient noise and nearby talkers on a face-to-face communication task," *J. Acoust. Soc. Am.*, vol. 34, pp. 936–941, 1962.

[9] A. Bradlow, G. Torretta, and D. Pisoni, "Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics," *Speech Communication*, vol. 20, pp. 255–272, 1996.

[10] J. G. Kidd, C. Mason, P. Deliwala, W. Woods, and H. Colburn, "Reducing informational masking by sound segregation," *J. Acoust. Soc. Am.*, vol. 95, pp. 3475–3480, 1994.

[11] P. Alku, J. Vintturi, and E. Vilkman, "Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation," *Speech Communication*, vol. 38, pp. 321–334, 2002.