Strategies adopted by talkers faced with fluctuating and competing speech maskers

Vincent Aubanel[a)] and Martin Cooke[b)]

Language and Speech Laboratory,

Universidad del País Vasco,

Paseo de la Universidad 5,

01006 Vitoria,

Spain

Abstract

Studying how interlocutors exchange information efficiently during conversations in less-than-ideal acoustic conditions promises to both further our understanding of links between perception and production and inform the design of human-computer dialogue systems. The current study explored how interlocutors' speech changes in the presence of fluctuating noise. Pairs of talkers were recorded while solving puzzles cooperatively in quiet and with modulated-noise or competing speech maskers whose silent intervals were manipulated to produce either temporally-sparse or dense maskers. Talkers responded to masked conditions by reducing the proportion of their speech which was in temporal overlap with the maskers, with larger relative reductions for sparse maskers. An analysis of talker activity in the vicinity of masker onset and offset events showed a significant reduction in onsets following masker onsets, and a similar increase in onsets following masker offsets. These findings demonstrate that talkers are sensitive to masking noise and respond to its fluctuations by adopting a "wait-and-talk" strategy.

PACS numbers: 4366Dc, 4370Bk, 4371Sy, 4372Dv

I. INTRODUCTION

Conversing in a noisy environment can be challenging for interlocutors engaged in a dialogue. A listener must solve the problem of understanding their partner's message against a background of other sound sources, while a talker may find the monitoring of their own utterances disrupted. The presence of other sound sources is known to affect both the way a talker speaks (e.g., Lombard, 1911; Dreher and O'Neill, 1957; Van Summers et al., 1988) and the information available to enable a listener to understand the message (e.g., see reviews in Pisoni and Remez, 2005). However, relatively little is known about whether interlocutors engage in behaviour aimed at actively reducing the impact of noise on dialogue, and, if so, what strategies are employed.

Compared to studies of the perceptual consequences of noise, most published work on the effect of noise on speech production has used relatively simple maskers such as white noise (e.g., Dreher and O'Neill, 1957; Van Summers et al., 1988) or fluctuating but temporally-dense stimuli such as speech babble (Pittman and Wiley, 2001; Garnier et al., 2010). Both are principally energetic maskers whose effect is to reduce the amount of undistorted speech information for talkers and listeners. Few production studies have examined noise types such as competing speech which in addition to their energetic masking effect also have an informational masking component (Carhart et al., 1969; Brungart, 2001). Competing speech poses additional problems for interlocutors since the listener may be uncertain as to which parts of the message belong to their interlocutor and which emanate from the background source, an effect which increases with target-masker similarity (Brungart, 2001; Vestergaard et al., 2009; Durlach et al., 2003). Talkers may also suffer in attempting to monitor their own voice when other speech material is present. Since natural dialogues typically contain significant amounts of overlapped speech (e.g., Jefferson, 1973; Kurtic,

[a)]Electronic address: v.aubanel@laslab.org; Also at Ikerbasque (Basque Science Foundation).
[b)]Also at Ikerbasque (Basque Science Foundation).

2012), both interlocutors face the additional problem of processing simultaneous talk from their own conversation in the presence of competing speech.

While competing speech in the conversational background can be seen as a complicating factor, it might at the same time present opportunities for talkers to deploy strategies whose aim is to ameliorate the negative consequences of masking on a foreground conversation. Naturally-occurring speech is highly non-stationary at a range of time scales, from syllable-rate amplitude modulations (Miller et al., 1984) to intra-turn pauses and inter-turn gaps (Heldner and Edlund, 2010). By modifying speech timing, a talker might be able to ensure that less of his/her speech is subject to masking by the background source. While it is difficult to envisage talker strategies which respond to syllable-scale intensity modulations in competing speech, it is conceivable that talkers are capable of exploiting pauses in the background source.

A few speech production studies (e.g., Webster and Klumpp, 1962; Lu and Cooke, 2008; Cooke and Lu, 2010) have examined the effects of competing speech on speech production. Webster and Klumpp (1962) found that competing speech induced a decrease in speech rate and an increase in communication errors for interlocutors engaged in a simple task involving communicating word lists. More recently, Lu and Cooke (2008) reported an increase in false starts and short pauses when reading sentence material in the presence of a competing talker. One specific possibility is that talkers reduce the proportion of their talk which overlaps with the background. Cooke and Lu (2010) tested this hypothesis by asking interlocutors to engage in a co-operative problem-solving task in quiet and in two types of fluctuating noise designed to possess similar amounts of energetic masking but differing in their degree of informational masking, viz. competing speech and speech-shaped noise modulated by the temporal envelope of speech. They found that talkers did indeed reduce the relative amount of speech in overlap with fluctuating maskers compared to a baseline in which speech elicited in quiet conditions was overlapped with the masker, with a somewhat greater reduction in the competing speech case (11 vs. 6 percentage points against a baseline of 40% overlap for quiet). Cooke and Lu (2010) speculated that the intelligibility of the competing

speech masker permitted greater overlap reduction through prediction of the likely timing of upcoming pauses in the masker. However, no explanatory basis for pause reduction in the unintelligible masker condition was provided.

Two issues motivated the current study. First, it is not clear what strategies talkers might use to reduce overlap with fluctuating maskers. Do talkers respond to masker onsets by curtailing their own speech activity, or do they await masker offsets before starting their contribution? Do talkers show any evidence of turn-taking with the background noise in the same way as is typical with their interlocutor? A second issue is the extent to which talkers adopt different strategies in the face of intelligible and non-intelligible maskers. Does competing speech allow better predictions of suitable times to speak? To address these questions, the current study adopted the cooperative task and the two masker types – competing speech and speech-modulated noise – used in Cooke and Lu (2010), but manipulated the length distribution of pauses in the masker to produce *dense* and *sparse* maskers, the former having pause lengths half that of the latter. We hypothesise that sparse maskers, which contain longer silent epochs, are more likely to give rise to effective active strategies which lead to a reduction in overlap.

Section II describes the maskers and corpus collection. Section III examines global effects of noise effects on speech production for the new corpus, and goes on to report the degree of overlap reduction exhibited by talkers. Section III also describes how onset and offsets in a talker's speech are affected by masker events. A new representation inspired by sensory neuroscience – the event-related activity plot – is presented which represents the onset and offset behaviour of talkers in the vicinity of masker changes, designed to pin down which specific speaker actions are responsible for overlap reduction.

II. CORPUS DESIGN AND COLLECTION

A. Overview

Speech material was elicited from pairs of talkers while they co-operated on a simple puzzle-solving task. Talkers participated in five conditions, one in quiet and four which contained masking noise throughout. The four masking conditions resulted from the combination of two masker types, competing speech (CS) and speech-modulated noise (SMN), with two pause length distributions which resulted in sparse and dense maskers.

B. Task

As in Cooke and Lu (2010), participants worked together on the solution of Sudoku puzzles. This is a natural and familiar task which elicits spontaneous speech, and which is capable of maintaining participants' interest over several sessions. The task needs little or no explanation and has a clear goal. Pilot studies revealed that puzzles graded 'easy' produced the greatest amount of interaction, largely by reducing the number of long pauses where talkers considered the next portion of the grid to tackle. Participants were provided with identical Sudoku grids, downloaded from the Daily Sudoku website[1] and were instructed to solve the puzzle jointly with their partner by writing digits on the supplied paper sheets. Participants were allowed to complete up to two puzzles in the imparted time; none of the pairs exceeded this limit.

C. Participants

Five pairs of native British English female speakers participated in the study (mean age: 20.1 years, SD=3.3), recruited through the University of Edinburgh Student and Graduate Employment service. In order to promote a natural form of interaction, a condition of participation was that speakers responding to the advertisement in pairs already knew each other well. All participants were familiar with Sudoku puzzles. They provided written

consent and were paid for their participation. Data collection was approved under the University of Edinburgh Ethical Review Procedure.

D. Maskers

Competing speech and speech-modulated noise maskers were constructed from part of the speech material used in Cooke and Lu (2010) consisting of a 10 minute recording of a female speaker talking aloud while solving Sudoku puzzles. A competing talker of the same gender was chosen to increase the informational masking potential (Brungart, 2001), while the challenge of foreground-background segregation was expected to be further increased through the use of masker material drawn from the same Sudoku scenario as the foreground task.

The recording of the masker was edited manually to replace non-speech sounds such as breath noise, laughing and coughing by silence (see section II.B of Cooke and Lu, 2010, for details), resulting in an alternating sequence of intelligible speech and silent intervals. Prior to manipulation, silences made up 49.5% of the recording. To construct the sparse competing speech maskers, the durations of all silent intervals in the original recording were increased. Conversely, for the dense masker all pause durations were decreased. Expansion and contraction factors were chosen so that the mean pause length in the sparse condition was twice that of the dense condition. This process solely modifies the distribution of pause lengths, leaving the speech material intact. Sparse and dense speech-modulated noise maskers were generated by amplitude modulating speech-shaped noise with the short term envelope of the sparse and dense competing speech maskers, following the procedure described in Brungart (2001). The speech-shaped noise sample had the same long-term spectrum as the competing speech sample following pause removal. The duration of the manipulated recording following contraction of pauses was just over 500 s, so to prevent repetition of masked speech material all maskers were truncated to have a duration of exactly 500 s.

The four masker signals are denoted CS-DENSE, CS-SPARSE, SMN-DENSE, and SMN-

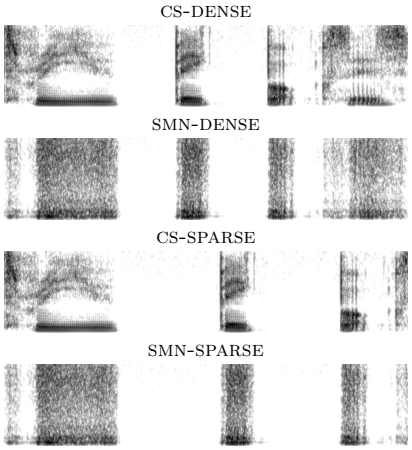


FIG. 1. Masker design. The speech portions are identical across dense and sparse maskers, while the silent intervals in dense maskers are half the duration of those for the sparse maskers. The utterance in CS-DENSE is *Three big boxes*.

SPARSE. Note that the procedure for generating the SMN pair of maskers ensures that the position and duration of pauses in CS-DENSE is the same as those of SMN-DENSE, with a similar identity for the sparse maskers. Figure 1 depicts fragments of the four maskers, aligned at the beginning of the same region of speech.

To avoid an effect of mean pause length on masker presentation level, versions of all 4 maskers without silent intervals were also generated. Masker level calibration for presentation was based on these pause-less signals: two adjustment values were established for CS and SMN pause-less signals to yield a value of 82 $\pm$0.1 dB SPL using a calibrated B&K 4100 HATS system. As in Cooke and Lu (2010), the level of 82 dB was chosen since it leads to a moderate-sized Lombard effect.

E. Procedure

Participants sat at either side of a table with a screen in the middle to prevent visual contact. In a similar setting, Fitzpatrick et al. (2011) found that the absence of visual information elicited a greater Lombard effect than was present when participants could see each

other. Participants wore Sennheiser ew 352 G3 head-mounted microphones and Sennheiser HD 650 open headphones throughout the recording, including in the QUIET condition. The headphones had a negligible attenuation effect on external sounds and therefore no own-voice feedback was felt to be necessary. Recordings were made using a custom PureData (Puckette, 2011) application which was also responsible for masker presentation.

After an initial training phase, pairs solved puzzles in quiet and in the four masked conditions, assigned in a Latin square order to provide counterbalancing across pairs. Each session lasted for 500 seconds, long enough to make good progress towards completion of one or more Sudoku puzzles.

F. Data analysis

In order to assess the degree to which maskers produced the kinds of robust effects observed in previous Lombard speech studies, $F0$ and $F1$ estimates were obtained every 10 ms during voiced epochs using PRAAT (Boersma and Weenink, 2012), which was also used to compute an energy contour. Speech rate was estimated as the number of syllable nuclei per second based on a syllable segmentation obtained from the prosogram (Mertens, 2004). For all four parameters, per-speaker median values averaged over all speakers are reported.

To support the main analysis of overlaps and foreground-background activity correlations reported in sections III.C and III.D below, speech was endpointed automatically for each channel using the endpointing module of the CMUSphinx recognition toolkit (Walker et al., 2004) and subsequently manually checked and corrected. The resulting segmentation had a temporal resolution of 10 ms. Silences, breath and other non-speech noises were marked as silence while speech and laughter were marked as speech.

Speech activity is defined as the binary-valued function of time $a_c(t)$ for talker or masker channel c which takes the value 1 if the frame contains activity and zero otherwise. Overlap between two channels c_1 and c_2 (i.e., pairs of talkers, or between a talker and masker) in the

time interval $[1, T]$ is then defined as

$$\Omega(c_1, c_2) = \frac{\sum_{t=1}^{T} a_{c_1}(t)\ a_{c_2}(t)}{\sum_{t=1}^{T} a_{c_1}(t)}. \quad (1)$$

Note that Ω is normalised by speech activity to cater for differences in activity both between talkers and conditions, and as a consequence is not commutative, i.e., $\Omega(c_1, c_2) \neq \Omega(c_2, c_1)$. To compute overlap values for the QUIET condition, where no masker was present, activity from the masked condition was substituted for a_{c_2} in eqn. 1 (see section III.C for further details of the overlap reference in the QUIET condition).

III. RESULTS

A. Performance on the task

On average, pairs filled in 50.1 digits in each condition, corresponding to 1.22 puzzles. The mean number of filled digits varied from 43.3 in CS-SPARSE to 55.1 in SMN-SPARSE, although conditions were statistically-equivalent [p=.12]. When normalised for differences in the amount of speech activity (see section III.C), pairs produced solutions at a rate which averaged 18.5 target digits per minute of speech. Again, while this quantity varied across conditions, with a low of 15.0 in CS-SPARSE to 20.4 in QUIET, large between-pair variability was evident, especially in the QUIET condition. Consequently, the target digit completion rate was marginally statistically-equivalent across conditions [$F(4, 36)$=2.4, p=.07, η^2=.21]. An analysis of the partially-completed puzzle sheets produced by each pair revealed an overall level of agreement on filled digits of 99.7%, demonstrating that participants found strategies which enabled them to transfer information successfully during the task.

B. Lombard effects

Figure 2 plots acoustic-phonetic parameter values in the five masking conditions. Compared to the no-masker condition, all maskers induced changes in the expected direction:

increases in speech energy, $F0$ and $F1$, and decreases in speech rate [all $p<.001$ apart from speech rate for CS-SPARSE, where $p<.05$].

Smaller increases in energy, $F0$ and $F1$, and larger decreases in speech rate are apparent for the SMN masker. However, separate two factor (noise type $\times$ masker sparsity) repeated-measures ANOVAs applied to the change in each parameter relative to the QUIET condition found no statistically-significant effect of noise type. A similar magnitude of effects between the two masker types is expected, echoing Cooke and Lu (2010), since both maskers were designed to produce a similar amount of energetic masking.

Relative to the quiet baseline, significantly smaller changes are observed in sparse maskers for $F0$ [$F(1,9)=8.1$, $p<.05$, $\eta^2=.13$] and $F1$ [$F(1,9)=16.6$, $p<.01$, $\eta^2=.30$], and a similar but not statistically-significant tendency can also be seen in the speech rate data. However, intriguingly, dense maskers did not induce larger changes in talkers' speech output level than sparse maskers.

To further understand this result, Figure 3 shows talker energies computed independently for three states: (i) epochs where no other noise was present (i.e., neither the masker nor interlocutor were active); (ii) epochs where the talker overlapped with the masker only (i.e., the interlocutor was not active) and; (iii) epochs where the talker overlapped with the interlocutor only (i.e., the masker was not active). Speakers did increase output level when the masker was active, but by only 0.74 dB on average, which is far less than the 6 dB difference between QUIET and the average of the masker conditions. It appears that talking in the presence of a fluctuating masker leads to an energetic Lombard effect even for those epochs when the masker is not active. A second feature of this data is the moderate *reduction* of around 1.5 dB in output level when interlocutors overlapped with each other in the absence of masker activity. Again, even though the masker was not active at these points, a masked-condition effect can be seen: the interlocutor overlap energy reduction is larger than in the equivalent no-masker condition.

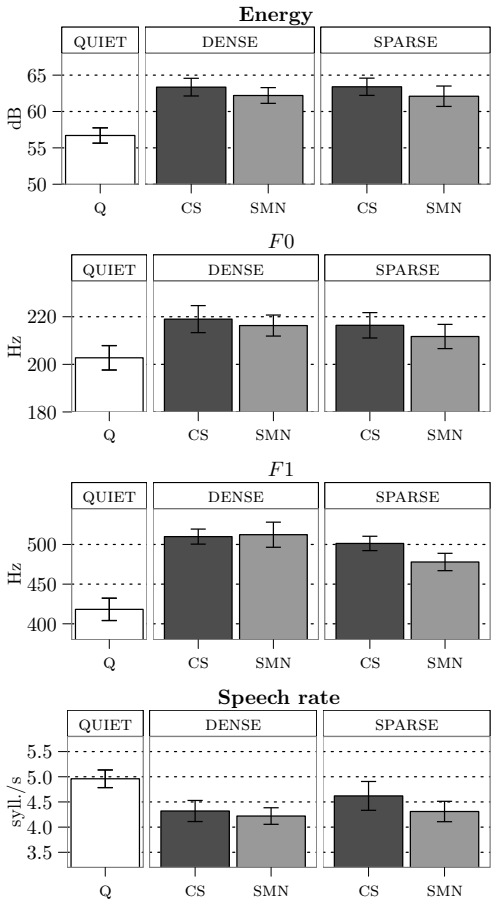


FIG. 2. Across-speaker means of energy, fundamental frequency, first formant frequency and speech rate. Error bars, here and elsewhere, represent ±1 standard error computed over the 10 speakers.

C. Talker-masker overlaps

Figure 4 shows mean speech activity and mean overlap $\Omega(s, m)$ between talker s and masker m, relative to the QUIET condition. While overlap is, of course, undefined in quiet due to the absence of the masker, a reference value can be obtained, as in Cooke and Lu (2010), from the overlap of speech activity in quiet and masker activity, which corresponds to the assumption of independence between talker and masker. Separate reference values for overlap are computed for the sparse and dense maskers.

One-way repeated-measures ANOVAs with noise type (Q, CS, SMN) as a within-subject

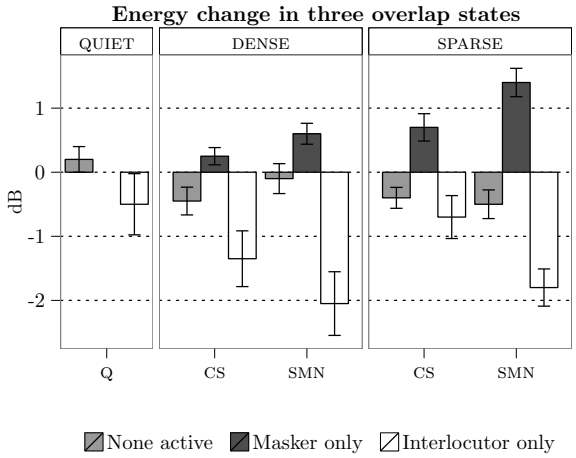


FIG. 3. Energy change for three states of overlap relative to the mean energy value reported in upper panel of Fig. 2

factor were carried out separately for dense and sparse maskers. A main effect of noise type was observed in both dense [$F(2,18)$=3.93, $p < 0.05$, η^2=0.3] and sparse [$F(2,18)$=7.37, $p < 0.01$, η^2=0.45] cases, and post-hoc pairwise comparisons, using Fisher's Least Significant Difference test, revealed that overlap in SMN-DENSE was significantly lower than in QUIET, and that both masker types were significantly lower than overlap in QUIET for the sparse maskers.

To enable comparison across these separate reference values, Figure 4 shows relative reductions in overlap, defined as the across-talker mean of the quantity

$$100\ \frac{\Omega_m(s,m) - \Omega_{quiet}(s,m)}{\Omega_{quiet}(s,m)}. \tag{2}$$

While sparse maskers tended to produce larger reductions in overlap, a two factor (noise type × masker sparsity) repeated-measures ANOVA on the relative overlap measure showed that reduction was independent of both noise type [$p = 0.51$] and masker sparsity [$p = 0.20$].

Further analysis reveals that overlap reduction is not primarily due to a decrease in raw overlap (i.e., the quantity expressed in the numerator of eqn. 1), which is statistically-equivalent across conditions when sparse and dense maskers are considered separately. In-

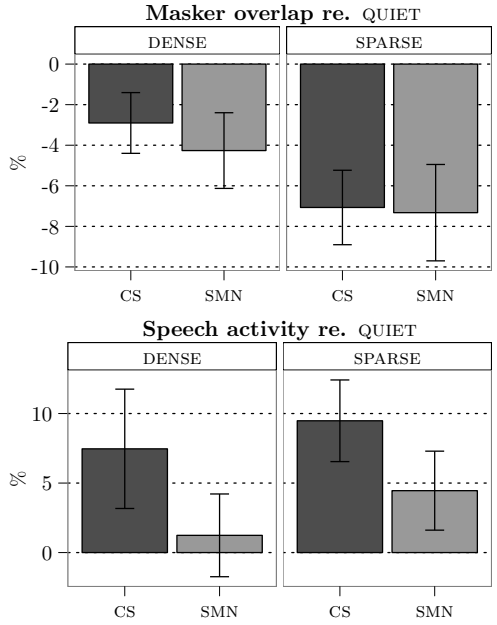


FIG. 4. Talker-masker overlap (upper) and talker activity (lower) relative to QUIET. Mean baselines over talkers for overlap in QUIET are 61.8% (dense) and 45.3% (sparse). Mean baseline for speech activity in QUIET is 32.1%.

stead, a large part of the reduction stems from an increase in speech activity in the masked conditions compared to the QUIET condition [mean: 5.7%, $t(39)$=3.39, p<0.01], as shown in the lower panel of Figure 4. It appears that talkers respond to maskers by increasing the amount of speech produced, yet maintaining a constant amount of overlap with the masker, resulting in a net *decrease* of overlap relative to speech activity.

D. Event-related activity

While previous sections have described gross overlap statistics, we now examine in more detail the influence of masker activity on talker activity by averaging the time course of speech activity in the vicinity of events – onsets and offsets – in the masker signal or the interlocutor's speech. We adopt a method to relate events to activity inspired by the reverse correlation technique (de Boer and Kuyper, 1968; Ringach and Shapley, 2004). Reverse correlation treats each observed output event as the consequence of processing a known

input sequence through an unknown system. An estimate of the influence of the system is then derived by averaging, across all output events, the input sequences which occur in their temporal vicinity.

Here, an *event-related activity* or ERA, ϵ, is computed by sampling speaker activity in a time window separately for all events of a given type, namely masker onsets and offsets and interlocutor onsets and offsets, as follows:

$$\epsilon(t) = \frac{\sum_{\tau \in E} a_c(t+\tau)\omega(t+\tau)}{\sum_{\tau \in E} \omega(t+\tau)},\ t \in [-T_1, T_2] \tag{3}$$

where E is the set of event times, T_1, T_2 are the limits of the window used to sample speech activity and ω is a binary weighting used to select the interval containing a *single* event centred at τ, since it is common for more than one event to be present in the window over which the ERA is computed. ϵ takes on values in the interval $[0, 1]$, where 0 indicates that no activity was observed in the corpus at a given time relative to an event type, and 1 means that activity was always observed at that point. Values in the range $0.15 - 0.45$ are typical. Events of a given type are pooled across talkers in the ERA calculation.

Figure 5 depicts ERA plots for speech activity around 4 types of event: interlocutor onsets and offsets and masker onsets and offsets. Consider first event-related activity for interlocutor events (upper plots). The left panel shows that, not surprisingly, speaker activity increases in the vicinity of an interlocutor offset, a consequence of turn-taking. What is interesting is that the increase in activity starts at least 1 second prior to the offset itself, and continues for about 0.5 s after the offset. Careful inspection reveals that the slope of the ERA curves is shallower for the interval preceding the actual offset. This difference in slope may be attributable to one of the "golden rules" of turn-taking (Sacks et al., 1974), that one of the most frequent behaviours at a turn change is latching, where speakers leave a small gap between the previous turn and their onset. The consequent concentration of onsets immediately following interlocutor offsets may give rise to an increase in slope of the ERA curve post-event.

The pattern for interlocutor offsets is reversed for interlocutor onsets (right panel).

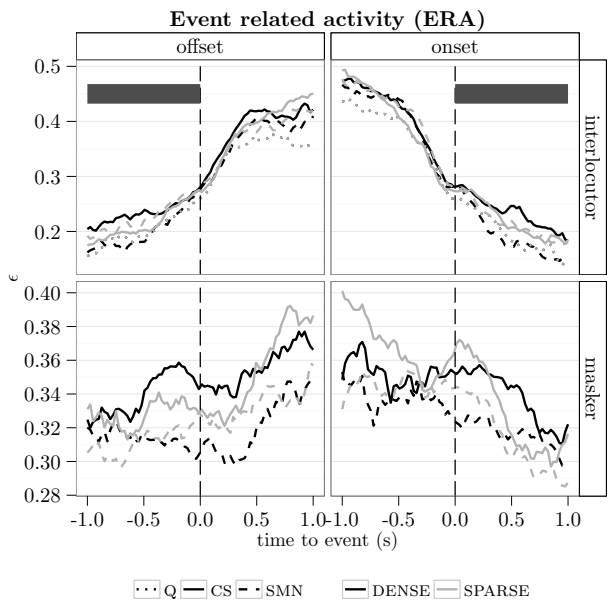


FIG. 5. ERA plots of speaker activity around masker and interlocutor onsets and offsets in the four masked conditions. Thick lines indicate background activity.

In fact, ERA plots for offset and onset have been drawn in this way to emphasise the effective continuity of the process from the left to the right panel: activity increases around interlocutor offsets, most likely as a response to interlocutor turn-yielding cues, and continues until the talker hands over the turn, in the vicinity of an interlocutor onset.

Comparing ERA plots for the QUIET and masked conditions, little influence of a noise background on interlocutor turn-taking can be seen apart from an overall increase in the baseline, reflecting the increased activity in masked conditions shown earlier in Figure 4.

We now turn to event-related activity around masker onsets and offsets (lower plots). If talkers were turn-taking with the masker the patterns here would be like those in the upper plots, which is clearly not the case. However, there is a very striking pattern of increased activity starting around 200−300 ms after masker offsets, and a similar decrease in activity at around the same point post-masker onset. Again, treating the left and right plots as representing the boundaries of an 'average' masker event (offset followed by onset), a rise followed by a fall in talker activity during the masker-free interval is evident. The

effect is much smaller than for interlocutor turn-taking: the activity axis has been expanded relative to the interlocutor case. On average, ϵ increases from around 0.3 to 0.35 for the SMN-DENSE masker, with a similar increase from a higher baseline for the other maskers. For interlocutor-based turn-taking, the change is about 4 times larger. Another difference between responses to interlocutor and masker events is in the predictive component. While there is some variation across masker types and densities, in general the bulk of activity change takes place several 100 milliseconds post-event, emphasising that speakers are responding to concrete events in the background but not, on this evidence at least, predicting them. The possible exception is for the competing speech maskers, particularly in the sparse case, where some anticipatory talker activity change is visible.

The differences in interlocutor and masker turn-taking behaviour are more apparent if onset and offset ERA plots are combined into a single contrast function

$$d = \epsilon_{\mathrm{ON}} - \epsilon_{\mathrm{OFF}} \tag{4}$$

where ϵ_{ON} and ϵ_{OFF} denote ERA functions for onsets and offsets respectively. The resulting curves are shown in Figure 6. Note that the zero-point in time now represents an event, and does not distinguish between onsets or offsets. These curves emphasise differences between what is happening pre- and post-events. The shoulder in activity at around $+200{-}300$ ms for the masker curves is evident.

ERA plots are useful in summarising turn-taking (with and without maskers), but they do not explicitly identify what it is that the speaker is doing which results in the activity change. Consider possible responses to a masker onset. The activity decrease observed in this case can be a consequence of a speaker refraining from starting to speak in the interval (a WAIT strategy), or it can be due to a greater tendency for speakers to STOP in response to masker onsets, or to the cumulative effect of both strategies. Similarly, speakers might react to a masker offset by starting to speak – TALK– or they might react by continuing to speak (CARRY ON). Figure 7 schematises these 4 strategies. Note that while each of these strategies has its reciprocal which results in the opposite effect in speech activity, we have

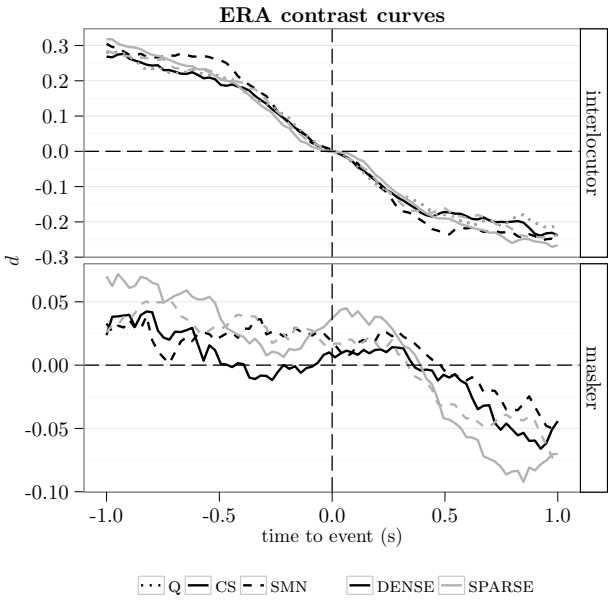


FIG. 6. Contrast curves (d in eqn. 4) for the four masked conditions.

only highlighted those strategies which appear potentially beneficial to talkers.

To seek evidence for each of these putative strategies, we compared counts of speaker events (separately for onsets and offsets) in the vicinity of masker events (again, separating onsets or offsets) in two intervals, one ending immediately prior to the event, the other starting at a fixed delay after the event. This computation used an interval of 300 ms for accumulating events, and a post-event delay of 200 ms, the latter chosen based on the minimum voicing time (e.g., Izdebski and Shipp, 1978) and stopping time (e.g., Ladefoged et al., 1973) in reaction to a signal. Variation of duration intervals from 200 to 500 ms and delays from 100 to 300 ms gave similar patterns of response. Table I presents results of χ^2 tests performed on the counts of the two time intervals used for comparison.

Table I demonstrates that the greatest change in speaker activity is in speaker onsets, in the vicinity of both masker offsets, where they show an increase, and masker onsets, where fewer speaker onsets occur. This provides clear evidence for the TALK and WAIT strategies. While speaker offset changes are less consistent, for the case of the CS-SPARSE masker there is evidence of a STOP strategy at work, with around 20% more offsets following masker

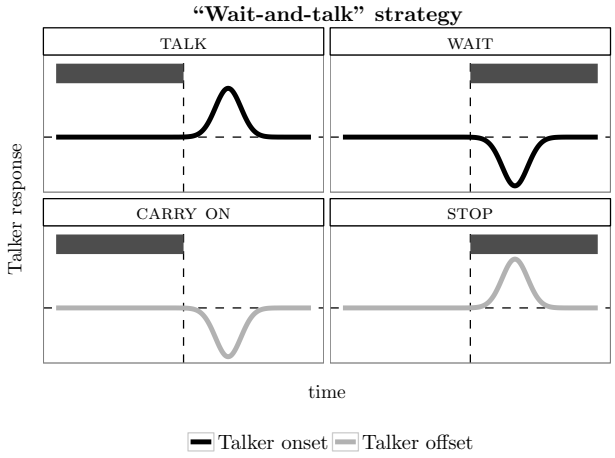


FIG. 7. Four hypothetical strategies which a talker might use in response to masker events. Thick lines depict masker activity (offsets in left column, onsets in right column) while curves schematise the likelihood of talker response (onsets in top row, offsets in bottom row). Vertical dashed lines indicate the masker event origin, while horizontal dashed lines represent the baseline of talker responses.

onsets than preceding. In fact, consistent differences across the four types of masker are observed, with the CS-SPARSE masker leading to the largest changes in speaker activity, and SMN-DENSE generally showing the smallest effects.

IV. DISCUSSION

A. Global effects of fluctuating maskers

Talkers faced with fluctuating maskers made global changes to speech parameters such as output level, fundamental frequency and speech rate consistent with previous descriptions of the Lombard effect (e.g., Van Summers et al., 1988). However, the intermittent nature of the masking noise used in the current study permitted further facets of talkers' responses to emerge. First, the commonly-observed increase in speech output intensity in noise appears on closer inspection to have two components. The first increment, which at less than 1 dB is

rather modest, is seen when the masker co-occurs with the talker's speech. The second and much larger increase in production level comes from noise being present in the experimental condition, regardless of whether the masker is active at any given moment. This finding calls into question the idea that talkers increase their level as an immediate 'reflex' in the face of noise (Lombard, 1911; Pick et al., 1989). Instead, the conditional awareness of noise being present seems to drive a global increase in level. An alternative interpretation is that a talker's compensatory response to noise is immediate, but requires time to return to its pre-compensatory level, just as is the case for a talker's response to formant frequency perturbation (Purcell and Munhall, 2006). Even the sparse stimuli used here may not have provided sufficiently-long masker-free intervals for the restoratory process to happen. Further studies will be required to clarify this issue.

A second finding is that, on average, interlocutors increased their speech level to a lesser extent when in overlap with each other during those epochs when the masker was not active. Part of this effect is likely to stem from energy transients related to one member of the pair starting to speak while the other stops. However, the fact that the effect is largest in less ideal conditions (e.g., SMN-DENSE) compared to QUIET or CS-SPARSE (where talkers were the most reactive to temporal changes in the masker, as evidenced by Table I) might suggest that adverse conditions lead to an increased monitoring of interlocutor's speech. While in ideal conditions, or when it appears to be possible to overcome disruption by exploiting the silences in the masker, talkers may allow for loss of information, when message reception is more difficult, talkers may additionally need to listen to the information present in overlaps, and therefore lower their own speech level. A similar pattern of differential energy increase was found by Aubanel et al. (In press) in a conversational setting, where talkers increased speech output to a greater extent during background overlaps as compared to interlocutor overlaps.

B. Overlap reduction

Talkers reduced the proportion of their speech which was in overlap with maskers. Qualitatively, this outcome confirms the findings reported by Cooke and Lu (2010), although in that study talkers managed to reduce overlap by 14-25%, substantially more than the 3-8% in the current study. Talkers in Cooke and Lu (2010) also achieved a larger reduction in overlap for competing speech than for modulated noise maskers, while here the reduction was similar for the two types of masker. This discrepancy may have been due to the possibility that talkers in the current study were less able to predict upcoming gaps in the masker due to the manipulation of pause lengths used to create sparse and dense maskers. Such disruptions would be expected to have most effect on the competing speech, and might account for the smaller degree of overlap reduction observed. Another possibility is that talkers in the two studies differed in their level of task engagement, perhaps due to differences in factors such as instructional nuances. It is conceivable that even a slight change in emphasis could result in talkers adopting different tolerances to masker overlap, and it may have been that talkers in the current study were more focused on task completion than those in the earlier study. One relevant fact here is the counter-intuitive finding that overlap reduction was achieved in large part through an *increase* in speech activity, which is a normalising factor in the proportion of speech in overlap (eqn. 1). An increase in activity in the presence of a masker can arise if talkers feel under pressure to complete as much of the task as possible, and is compatible with the adoption of a "wait-and-talk" strategy as discussed further below. Under a less overt task imperative, talkers might be expected to maintain or even reduce the amount of speech activity. In support of this notion, a study of the influence of live background conversations on foreground dialogues in the absence of a task (Aubanel et al., In press) found no increase in speech activity when measured relative to a no-background condition.

C. Talker strategies to cope with fluctuating noise

The current study explored the strategies that talkers use to reduce overlap. The evidence presented suggests that when confronted with fluctuating noise, talkers are able to retime their onsets and offsets in response to changes in masker activity. These adjustments amount to a weak form of turn-taking with the noise but with a key difference compared to normal (i.e., interlocutor) turn-taking. While talkers are capable of timing their onsets to match the offsets of their interlocutor, and even precede them by as much as 500 ms, there is an almost complete absence of this predictive component for the case of timing with respect to masker events. Instead, talker behaviour in response to maskers is best-described as reactive, with a delay of around 300 ms in evidence. Of course, it is difficult to envisage the basis for predicting modulated masker onset and offset events, but there is a hint of predictive capacity for the sparse competing speech masker. As suggested above, the manipulation of natural pause lengths might have disrupted listeners' ability to identify upcoming masker events.

In controlled experimental settings, talkers have been shown to require a minimum of 200 ms in reaction to a signal to produce a vocal sound (Fry, 1975; Izdebski and Shipp, 1978; Shipp et al., 1984) and can stop with the same delay (Ladefoged et al., 1973). These values are slightly shorter than the delays we observed in starting and stopping in response to masker offsets and onsets, a difference which might be due to the greater cognitive complexity of our task which required monitoring interlocutor's speech as well as the masker while problem-solving, and also because talkers were not instructed to respond to these specific signals. Models proposed to account for how talkers stop speaking can be divided in two different classes. One set posits an immediate reaction to the signal to stop (Nooteboom, 1980; Hartsuiker and Kolk, 2001) in reaction to extrinsic factors. When linguistic processing is required to establish the need to stop, reaction time increases. For example, it took 300 ms for talkers to stop speaking when they needed to evaluate the meaning of a word used as a stop signal in a picture naming task (Slevc and Ferreira, 2006), and this delay

exceeded 400 ms in a task where talkers had to interrupt the naming of a picture which was changed on a small proportion of trials (Hartsuiker et al., 2008). The other class of models invoke intrinsic factors such as self-monitoring, and considers the possibility that speakers strategically plan their action. Levelt (1983) showed that upon detection, talkers avoid halting word-internally when the current word is correct, as a way to signal to the addressee that only interrupted words were erroneous (see also Levelt, 1989). More recently, Seyfeddinipur et al. (2008) found that speakers can postpone their interruption until resumption is possible, in a way to avoid sounding disfluent to their interlocutor or losing the turn.

While the data presented here do not allow direct testing of the hypothesis that talkers were postponing their activity upon perceiving a masker change, the presence of a shoulder at 300 ms in Fig. 6 suggests that speakers were reacting to the masker as soon as they could. In the current setting, unlike in the experimental studies outlined above, speakers were not prompted explicitly to react to the masker change. Instead, they adopted this behaviour in order to complete the task through conversation. The idea that talkers act as soon as the need to stop or start is detected as opposed to strategic planning is further reinforced by the pattern of results seen for the four active behaviours which together make up the "wait-and-talk" strategy. The easiest case for talkers, and the one for which there is the clearest evidence in our data, is the WAIT action: in this case, speakers only have to postpone their already planned speech, e.g., until the masker stops, without necessarily having to interrupt their speech. In TALK, the converse can be envisaged, where speakers resume their postponed speech following a release from the masker. The STOP action however requires an interruption of speech, and is only observed for the CS-SPARSE case. The more difficult case to implement is that of CARRY ON, where speakers have to plan and execute new speech for taking advantage of the extra opportunity offered by the release of the masker. It seems unlikely that speakers can achieve this behaviour in the limited time windows afforded by maskers with pause distributions typical of natural speech. Also, the data gathered in the CARRY ON situation target cases in which speakers are mostly already overlapping with the masker, possibly capturing those turn parts for which timing is more dependent on the needs

of the interaction. For example, Heldner et al. (2011) found that 37% of very short utterances (shorter than 1 s) occurred less than 200 ms following interlocutor's speech, therefore the retiming of this class of utterances might be limited in adjustment to fluctuating background noise.

Stronger "wait-and-talk" effects were obtained for competing speech, and in particular in the CS-SPARSE condition. As noted above, listeners (who are also talkers) might be capable of predicting background events to some degree for intelligible maskers. The informational content of intelligible maskers would thus be beneficial to talkers in this natural scenario, pointing to a more flexible use of background information than is usually found in studies using more controlled settings (e.g. Mattys et al., 2009). Another possibility is that because of the added disruption arising from the informational masking component of competing speech (Carhart et al., 1969; Brungart, 2001), speakers might feel more of a need to implement strategies to limit the effect of the noise. As Hazan and Baker (2011) found, production changes in clear speech appear to be tailored to the communication barrier that the listener is perceived to be confronted with, and are of a smaller order of magnitude when engaged in a real communicative situation than when asked to speak clearly, at which points talkers position their speech at a different point on the hyper-hypo continuum (Lindblom, 1990; Moore and Nicolao, 2011).

While this study shows that a wealth of information can be obtained from voice activity only, further insights might come from examining the interactional function of those speech parts – e.g., initiations or truncations – which are found in reaction to changes in masker activity. For applications such as automatic dialogue systems operating in noisy environments, it will be useful to know not only where human talkers interrupt their speech in terms of the syntactic structure of the message (e.g., Tydgat et al., 2011), but also what are the most appropriate interactional locations for dealing with the adverse condition while remaining intelligible to the user.

V. CONCLUSIONS

Talkers react to fluctuating maskers in ways which are only partially-described by classical Lombard effects. In a task involving cooperative problem solving in quiet and noise, increases in energy commonly observed in the presence of noise compared to speech in quiet occurred in conditions where maskers were present, regardless of whether the masker was active in any given epoch. Talkers reduced the amount of their speech in overlap with a masker, and showed clear effects in response to masker onsets and offsets. Talkers adopted a "wait-and-talk" strategy, reducing the likelihood of onsetting following masker onsets, and increasing onset activity subsequent to masker offsets.

Acknowledgements

This work was supported by the LISTA Project (`http://listening-talker.org`), funded from the Future and Emerging Technologies programme within the 7th Framework Programme for Research of the European Commission, FET-Open grant number 256230. We thank Julián Villegas for providing the PureData patch used for stimuli presentation and recordings and Rob Clark for help during data collection at the University of Edinburgh.

Endnotes

1. http://www.dailysudoku.com (Last viewed July 10, 2012).

References

Aubanel, V., Cooke, M., Foster, E., García Lecumberri, M. L., Mayo, C., In press (2012). Effects of the availability of visual information and presence of competing conversations on speech production, in: Interspeech, Portland, US.

de Boer, E., Kuyper, P., 1968. Triggered correlation. IEEE Trans. Biomed. Eng. 15, 169–179.

Boersma, P., Weenink, D., 2012. Praat: doing phonetics by computer [Computer Program]. Version 5.3.11. http://www.praat.org (Last viewed July 10, 2012).

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., Tillman, T., Greetis, E., 1969. Perceptual masking in multiple sound backgrounds. J. Acoust. Soc. Am. 45, 694–703.

Cooke, M., Lu, Y., 2010. Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. J. Acoust. Soc. Am. 128, 2059–2069.

Dreher, J. J., O'Neill, J. J., 1957. Effects of ambient noise on speaker intelligibility for words and phrases. J. Acoust. Soc. Am. 29, 1320–1323.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., Kidd Jr., G., 2003. Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. J. Acoust. Soc. Am. 114, 368–379.

Fitzpatrick, M., Kim, J., Davis, C., 2011. The effect of seeing the interlocutor on speech production in different noise types, in: Interspeech, Florence, Italy. pp. 2829–2832.

Fry, D. B., 1975. Simple reaction-times to speech and non-speech stimuli. Cortex 11, 355–360.

Garnier, M., Henrich, N., Dubois, D., 2010. Influence of sound immersion and communicative interaction on the Lombard effect. J. Speech. Lang. Hear. R. 53, 588–608.

Hartsuiker, R. J., Kolk, H. J., 2001. Error monitoring in speech production: A computational test of the perceptual loop theory. Cog. Psych. 42, 113–157.

Hartsuiker, R. J., Catchpole, C. M., de Jong, N. H., Pickering, M. J., 2008. Concurrent processing of words and their replacements during speech. Cognition 108, 601–607.

Hazan, V., Baker, R., 2011. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. J. Acoust. Soc. Am. 130, 2139–2152.

Heldner, M., Edlund, J., 2010. Pauses, gaps and overlaps in conversations. J. Phon. 38, 555–568.

Heldner, M., Edlund, J., Hjalmarsson, A., Laskowski, K., 2011. Very short utterances and timing in turn-taking, in: Interspeech, Florence, Italy. pp. 2837–2840.

Izdebski, K., Shipp, T., 1978. Minimal reaction times for phonatory initiation. J. Speech Hear. Res. 21, 638–651.

Jefferson, G., 1973. A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences. Semiotica 9, 47–96.

Kurtic, E., 2012. Overlapping talk and turn competition in multi-party conversations. Ph.D. thesis. University of Sheffield. Sheffield, 271 p.

Ladefoged, P., Silverstein, R., Papçun, G., 1973. Interruptibility of speech. J. Acoust. Soc. Am. 54, 1105–1108.

Levelt, W. J. M., 1983. Monitoring and self-repair in speech. Cognition 14, 41–104.

Levelt, W. J. M., 1989. Speaking: From intention to articulation. MIT Press, Cambridge, MA, 566 p.

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

Lombard, E., 1911. Le signe d'élévation de la voix (The sign of the rise in the voice). Ann. Malad. Oreille, Larynx, Nez, Pharynx (Ann. Dis. Ear, Larynx, Nose, Pharynx) 37, 101–119.

Lu, Y., Cooke, M., 2008. Speech production modifications produced by competing talkers, babble, and stationary noise. J. Acoust. Soc. Am. 124, 3261–3275.

Mattys, S., Brooks, J., Cooke, M., 2009. Recognizing speech under a processing load: Dissociating energetic from informational factors. Cog. Psych. 59, 203–243.

Mertens, P., 2004. The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model, in: Speech Prosody, Nara, Japan. pp. 23–26.

Miller, J. L., Grosjean, F., Lomanto, C., 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. Phonetica 41, 215–225.

Moore, R. K., Nicolao, M., 2011. Reactive speech synthesis: Actively managing phonetic contrast along an H&H continuum, in: ICPhS, Hong Kong. pp. 1422–1425.

Nooteboom, G. S., 1980. Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech, in: Fromkin, V.A. (Ed.), Errors in linguistic performance. Academic Press, New York, pp. 87–95.

Pick, H. L., Jr., Siegel, G. M., Fox, P. W., Garber, S. R., Kearney, J.K., 1989. Inhibiting the Lombard effect. J. Acoust. Soc. Am. 85, 894–900.

Pisoni, D. B., Remez, R. E. (Eds.), 2005. The Handbook of Speech Perception. Blackwell, Malden, MA, 708 p.

Pittman, A. L., Wiley, T. L., 2001. Recognition of speech produced in noise. J. Speech. Lang. Hear. R. 44, 487–496.

Puckette, M., 2011. PureData [Computer Program]. Version 0.42-5. http://puredata.info (Last viewed July 10, 2012).

Purcell, D. W., Munhall, K. G., 2006. Compensation following real-time manipulation of formants in isolated vowels. J. Acoust. Soc. Am. 119, 2288–2297.

Ringach, D., Shapley, R., 2004. Reverse correlation in neurophysiology. Cognitive Sci. 28, 147–166.

Sacks, H., Schegloff, E. A., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735.

Seyfeddinipur, M., Kita, S., Indefrey, P., 2008. How speakers interrupt themselves in managing problems in speaking: Evidence from self-repairs. Cognition 108, 837–842.

Shipp, T., Izdebski, K., Morrissey, P., 1984. Physiologic stages of vocal reaction time. J. Speech Hear. Res. 27, 173–178.

Slevc, L. R., Ferreira, V. S., 2006. Halting in single word production: A test of the perceptual loop theory of speech monitoring. J. Mem. Lang. 54, 515–540.

Tydgat, I., Stevens, M., Hartsuiker, R. J., Pickering, M. J., 2011. Deciding where to stop speaking. J. Mem. Lang. 64, 359–380.

Van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., Stokes, M. A., 1988. Effects of noise on speech production: Acoustic and perceptual analyses. J. Acoust. Soc. Am. 84, 917–928.

Vestergaard, M. D., Fyson, N. R. C., Patterson, R. D., 2009. The interaction of vocal characteristics and audibility in the recognition of concurrent syllables. J. Acoust. Soc. Am. 125, 1114–1124.

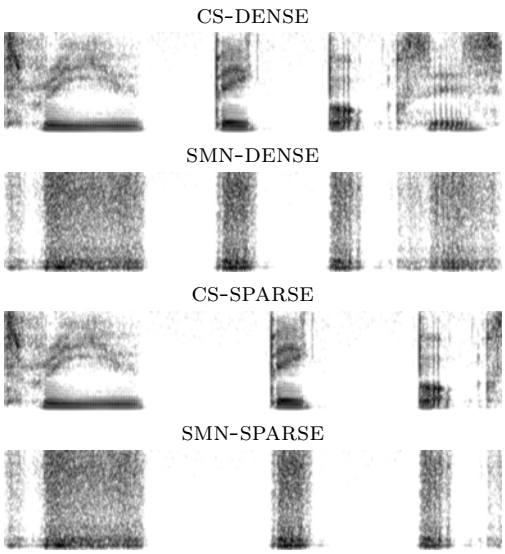
Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, J., 2004. Sphinx-4: A flexible open source framework for speech recognition. Technical Report TR-2004-139. Sun Microsystems, Inc. Mountain View, CA, 15 p.

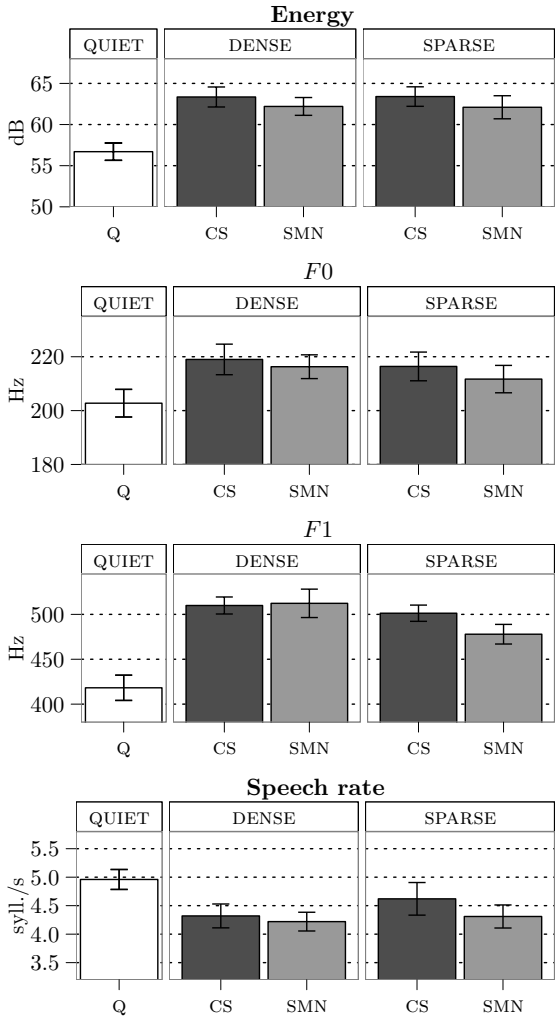
Webster, J. C., Klumpp, R. G., 1962. Effects of ambient noise and nearby talkers on a face-to-face communication task. J. Acoust. Soc. Am. 34, 936–941.

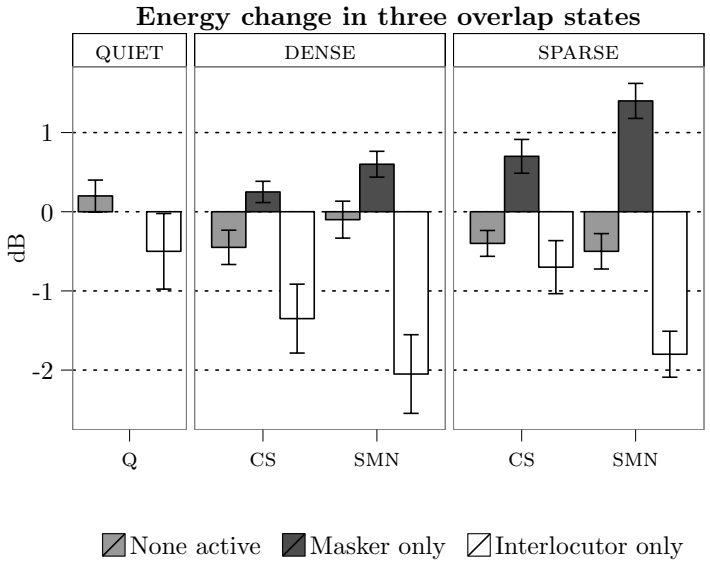
TABLE I. Counts of speaker events in the vicinity of masker events. N is the count of speaker events occurring in the interval $[-300; 0]$ ms relative to the masker event; % change is computed between this number and the counts of events occurring in the interval $[200; 500]$ ms following masker event (not shown). χ^2 tests are reported in the next two columns, comparing counts in the two intervals. Visual indication of statistical significance following classic p-values grouping is added in the rightmost column of each panel.

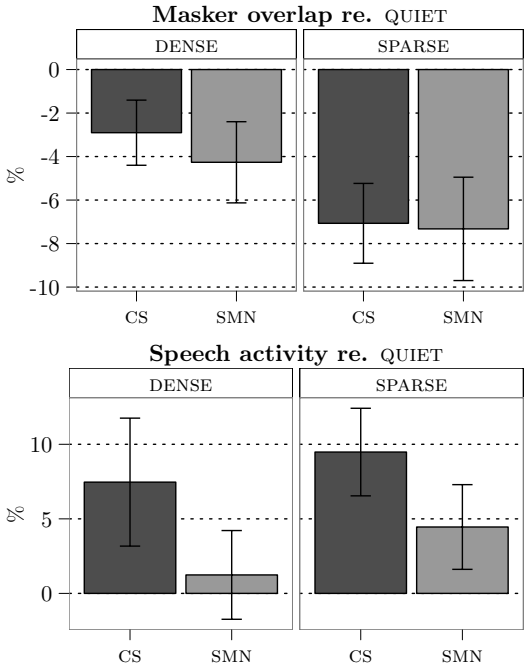
		Masker offset				Masker onset			
Speaker event	Condition	N	% change	χ^2 p	sig.	N	% change	χ^2 p	sig.
onset	CS-DENSE	241	21.7	5.14 $p < 0.05$	*	291	−20.3	6.68 $p < 0.01$	**
onset	SMN-DENSE	224	16.4	2.78 $p = 0.1$	.	231	6.6	0.48 $p = 0.49$	
onset	CS-SPARSE	175	31.2	7.37 $p < 0.01$	**	207	−35.5	15.87 $p < 0.001$	***
onset	SMN-SPARSE	162	25.0	4.50 $p < 0.05$	*	194	−25.4	7.17 $p < 0.01$	**
offset	CS-DENSE	261	5.9	0.45 $p = 0.5$		276	−8.4	1.02 $p = 0.31$	
offset	SMN-DENSE	238	−2.1	0.06 $p = 0.81$		245	11.0	1.41 $p = 0.23$	
offset	CS-SPARSE	185	−10.9	1.17 $p = 0.28$		172	21.8	3.71 $p = 0.05$	*
offset	SMN-SPARSE	143	13.9	1.29 $p = 0.26$		185	6.1	0.33 $p = 0.57$	

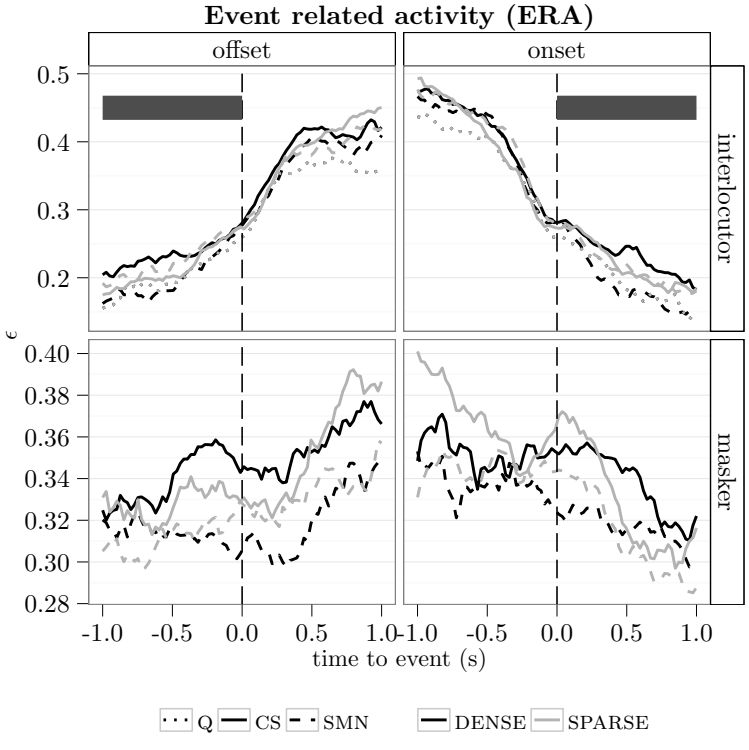
List of Figures

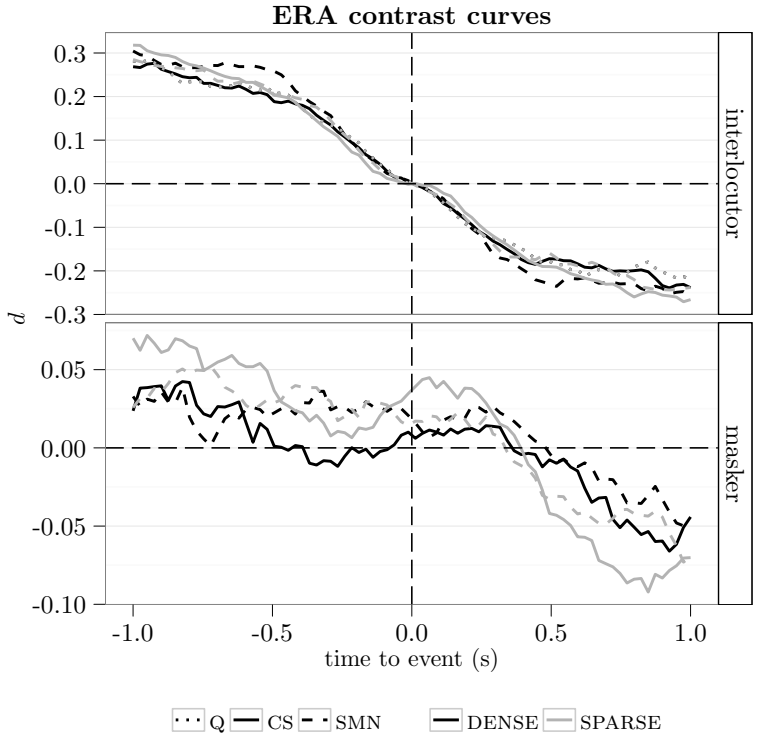

CS-DENSE
SMN-DENSE
CS-SPARSE
SMN-SPARSE

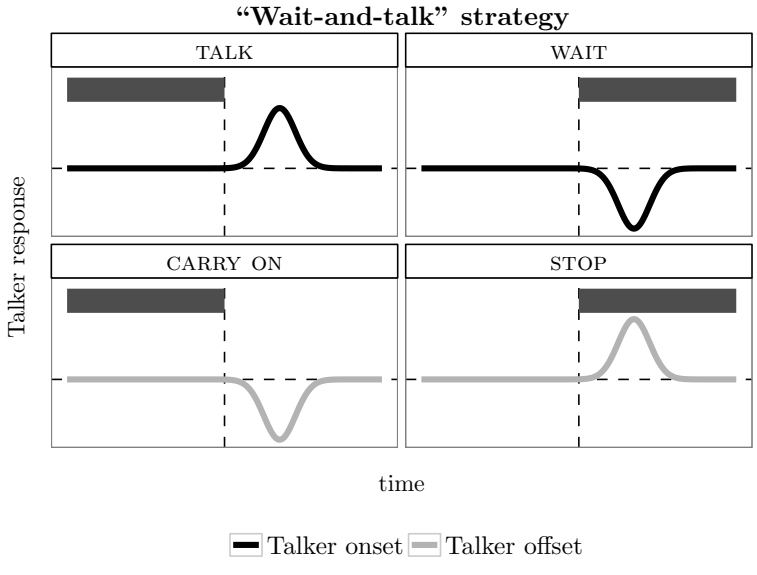

Energy
QUIET
DENSE
SPARSE
dB
65
60
55
50
Q
CS
SMN
CS
SMN
F0
Hz
220
200
180
F1
500
450
400
Speech rate
syll./s
5.5
5.0
4.5
4.0
3.5


Energy change in three overlap states
QUIET
DENSE
SPARSE
1
0
-1
-2
dB
Q
CS
SMN
CS
SMN
None active
Masker only
Interlocutor only


Masker overlap re. QUIET
DENSE
SPARSE
0
-2
-4
-6
-8
-10
%
CS
SMN
CS
SMN
Speech activity re. QUIET
DENSE
SPARSE
10
5
0
%
CS
SMN
CS
SMN


Event related activity (ERA)
offset
onset
interlocutor
masker
ϵ
0.5
0.4
0.3
0.2
0.40
0.38
0.36
0.34
0.32
0.30
0.28
-1.0
-0.5
0.0
0.5
1.0
time to event (s)
Q
CS
SMN
DENSE
SPARSE


ERA contrast curves
interlocutor
masker
d
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
0.05
0.00
-0.05
-0.10
-1.0
-0.5
0.0
0.5
1.0
time to event (s)
Q
CS
SMN
DENSE
SPARSE


"Wait-and-talk" strategy
TALK
WAIT
CARRY ON
STOP
Talker response
time
Talker onset
Talker offset