Effects of the availability of visual information and presence of competing conversations on speech production

Vincent Aubanel[1,2]*, Martin Cooke*[1,2]*, Emma Foster*[3]*, M Luisa Garcia Lecumberri*[1] *and Catherine Mayo*[4]

[1]Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
[2]Ikerbasque (Basque Foundation for Science)
[3]Department of Human Communication Sciences, University of Sheffield
[4]Centre for Speech Technology Research, University of Edinburgh

v.aubanel@laslab.org

Abstract

How do talkers maintain intelligibility when speaking in the presence of a background conversation? The current study identified acoustic and temporal modifications of speech manifested by interlocutors in the face of competing speech, with and without visual contact. Pairs of talkers held free conversations either alone or in the presence of a second pair. Regardless of the availability of visual information, speaking simultaneously with another talker resulted in overall increases in energy, F0, F1 and a decrease in speech rate. Overlapping with the background pair resulted in an increase in energy but no change in the two prosodic parameters F0 and speech rate. By contrast, within-pair overlap led to an increase in F0 and a decrease in rate, and no change in speech level. The absence of visual cues produced a significant reduction in within-pair overlap, which tended to be greater when the background pair was present. These findings emphasize the need to distinguish between Lombard and interactional influences on acoustic parameters, and suggest that adverse conditions such as competing speech or absence of visual cues cause interlocutors to adopt more careful dialogue strategies, perhaps with the aim of reducing energetic and informational masking at the ears of the listener.
Index Terms: simultaneous conversations, speech production modifications, speech in noise

1. Introduction

Speaking in a noisy environment usually prompts the talker to modify his or her speech in ways – collectively described as Lombard effects [1, 2] – which are well-understood, at least for stationary noise. Rather less is known about the common situation of talking in the presence of competing speech. In a communicative situation, competing speech not only disrupts message reception at a low-level by energetic masking, but can lead to potential confusions due to informational masking arising from factors such as similarity between audible components of target and masker speech [3, 4]. Discovering whether speakers deploy production strategies which seek to minimise the effects of both energetic and informational masking at the ears of their interlocutor is of interest both in understanding production-perception links, and for its potential application to context-aware speech output technology.

Of the few studies that have examined the influence of background on foreground speech, an early work found an increase in word pronunciation errors while communicating a word list to an interlocutor in the presence of a simultaneous pair engaged in the same task [5]. A recent study in a more natural setting discovered that competing speech led to more disfluencies and mistimings, and less rapid turn-taking, in pairs of speakers engaged in face-to-face conversations [6].

Being able to see the interlocutor could play a crucial role in this setting: in task oriented dialogue, speaking without visual contact with the interlocutor has been shown to result in longer task completion time, shorter turns, more frequent interruptions, more overlapping and a greater number of backchannels [7] as well as shorter inter-turn gaps [8]. When speaking in the presence of a variety of stationary and fluctuating background noises, speech output level was significantly higher in the absence of visual contact [9].

Here, the motivation for controlling visual contact is to induce contexts where speech modifications become necessary. By removing the visual channel, talkers are constrained to use exclusively auditory strategies to communicate. Further, when a visual modality condition is combined with the presence or absence of competing speech, a gradation of adversity is produced which might better identify the conditions under which speech modifications are most salient.

Section 2 describes the competing-conversations corpus and the methods employed to address these issues. Section 3 presents both Lombard effects and temporal overlap characteristics of speech produced in adverse conditions, highlighting the role of visual information.

2. Corpus and Methods

2.1. Corpus collection and annotation

Six pairs of British English female talkers were recorded while engaging in natural, unrestricted dialogues. Talkers were instructed to converse only with the other interlocutor in their pair. In each recording session, pairs were alone for half the time while both pairs were present for the other half. Pairs sat facing each other at a round table, so that when both pairs were present, talkers had to "talk across" the other pair. In half of the conditions, talkers wore shallow visors which prevented them from seeing their interlocutors but had no attenuating effect on audio transmission (Figure 1). Pair members knew each other but not the other pair, and their pairing remained the same in all experimental conditions. In total, the corpus contains 450 minutes of conversational speech.

Figure 1: *Recording set up, with speakers wearing visors.*

A skilled transcriber labelled turn constructional units (TCUs, see [10] adapted from [11]) with explicit coding of incomplete TCUs [−], incomplete words [⋆], elongations [:] and inbreaths [<] (see table 1). Turns were computed automatically in a post-processing step by merging adjacent TCUs or TCUs separated by a silence not exceeding 120 ms.

Annotation	*N*
turns	13259
TCUs	21700
including:	
incomplete TCUs	*6011*
incomplete words	*1750*
elongations	*2960*
inbreaths	*23*
words	85307
silences (> 120 ms)	12623
inbreaths	5270

Table 1: *Counts of annotation units*

The complete set of transcriptions amounts to 260 pages in a conversation analysis format. The excerpt in Figure 2 illustrates the nature of the speech and the turns/encoding conventions.

```
814.31  Ana:    oh ok
814.82  Annie:  cos they have to order in the
                vaccinations
815.99  Betty:  I know
816.29  Bea:    do you know what I mean it's a pure
                palava you have to get like work
                clo:thes. and all this kind of jazz it
                was just a-
816.54  Ana:    I might do that I kinda wanna go to the
                doctors anyway
817.31  Annie:  so i*-
819.98  Ana:    cos-
819.99  Annie:  but it's a nurse you have to see
820.71  Bea:    you had to worry about sales targets all
                that kinda- er- you know it wasn't as fun
821.33  Ana:    oh ok
822.89  Ana:    I keep getting erm:-
824.28  Betty:  yeah
824.36  Ana:    cramps in my legs
824.51  Bea:    whereas this like it's busy so you're
                kept moving you're not like-
826.05  Ana:    and my feet
826.97  Annie:  in the night
827.00  Betty:  mhm
827.74  Ana:    no like- at l*- like at least five ten
                times a day
828.29  Betty:  oh no that's really good becky I'm really
                happy for you
828.35  Bea:    yeah
829.13  Annie:  c*-
829.70  Bea:    yeah
```

Figure 2: *Conversation fragment. Conversing pairs are Ana/Annie and Bea/Betty (names were changed) during a no-visual-contact condition. Turn start times are listed in the first column.*

2.2. Analysis

The effects of noise on speech production usually include increases in energy, fundamental frequency ($F0$), frequency of the first formant ($F1$), and a decrease in speech rate [1, 2]. These four acoustic parameters were extracted from the epochs when speech was present. Energy, $F0$ and $F1$ were extracted with Praat [12] at 10 ms intervals. Measurement of the latter two was restricted to voiced segments, and both were transformed to cents relative to their median value in the entire corpus on a per-speaker basis. A proxy for speech rate was obtained using timings and pronunciations derived from the orthographic transcription, and expressed as the number of vowels per second in the canonical pronunciation as given by the BEEP dictionary [13]. In this coding, vowel counts correspond well to number of syllables in the word. The most frequent unrecognised words were added to the dictionary, which left 952 repetitions of 32 tokens occurring between 5 and 10 times and 493 tokens occurring 4 times or fewer. These unrecognized words were assigned a default value of 1 vowel. Interrupted words were ignored.

Lombard effects were computed both globally at the visual/non-visual condition level and at a finer-grained level, as a function of the number of simultaneous talkers, and further contrasting within- and across-pair overlaps. Within-pair overlap refers to those portions of speech where the background activity comes exclusively from the interlocutor's speech, while across-pair overlap denotes portions of speech when only the speaker and the background pair are simultaneously active. For these lat-

ter two analyses, values are computed as the difference of the parameter with the baseline condition where the speaker is the only speaker active. For each condition the overlap proportion was calculated separately for each speaker as the amount of overlapping speech divided by total speech activity for that talker.

3. Results

3.1. Lombard effects

$F0$, $F1$ and speech rate changed with the number of simultaneous talkers [all $p < .01$] in the expected direction as seen in earlier Lombard studies (Figure 3). Energy also increased relative to the single active talker condition, though not by much, and showed no further increase as additional talkers became active. However, although speech rate tended to be slower in the absence of visual contact, there was no significant difference between the visual contact and audio-only conditions for any of the parameters.

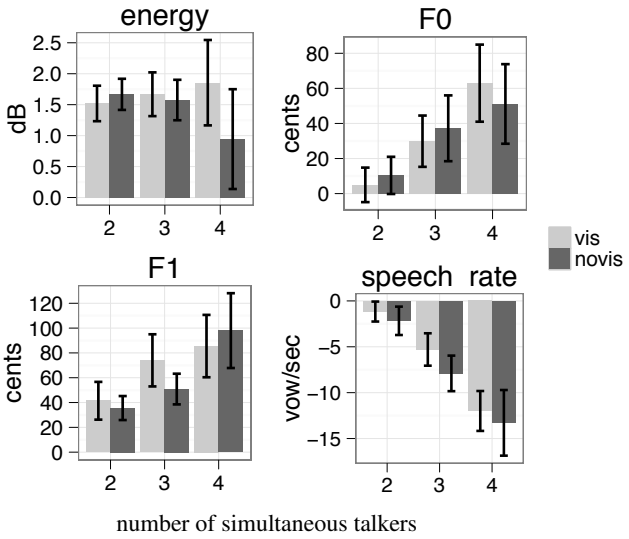

Figure 3: *Lombard effects as a function of the number of simultaneous talkers. Error bars, here and elsewhere, display +/- 1 standard error over all talkers ($N = 12$).*

3.2. Within- and across-pair overlaps

It is useful to distinguish overlapping speech within a single conversation from overlap across conversations. If purely Lombard effects were at work, then the presence of additional energy during speech production should result in similar effects in the two conditions. Figure 4 demonstrates that this is not the case. The prosodic parameters $F0$ and speech rate show much larger increases [$p < .001$] during within-pair overlaps, and no changes when occurring across pairs. In contrast, speech output level barely changes during overlap within a conversation but is clearly affected by an active background [$p < .01$]. Modality had no influence on any parameter apart from energy, where speakers spoke more quietly when conversing in the absence of visual contact [$p < .01$], although the reduction was small.

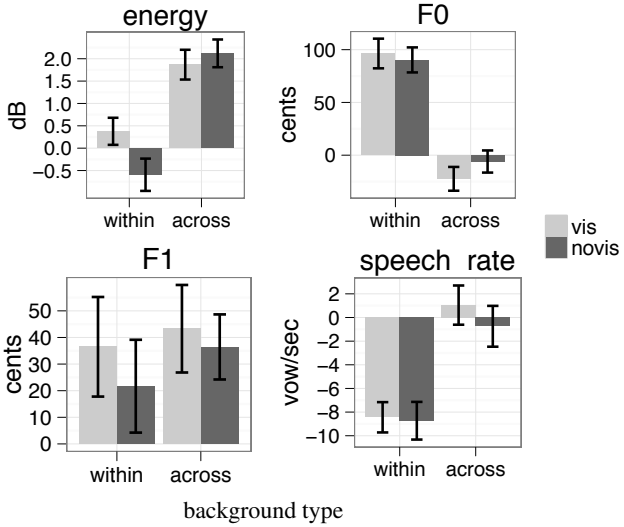

Figure 4: *Lombard effects contrasting within- and across-pair overlaps background types.*

3.3. Temporal overlaps

Conversational partners' speech was in overlap approximately 25% of the time in the absence of the other pair, and when visual contact was permitted. Removing visual contact led to a significant reduction (18%) in overlap [$p < .01$] while a marginal reduction of 9% occurred when the other pair were present [$p = .09$].

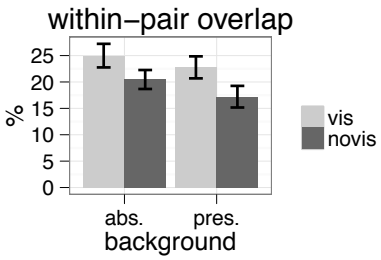

Figure 5: *Proportion of within-pair overlaps*

The across-pair overlap proportion (not shown) was high (across-talker mean: 80.0%) and statistically-equivalent in the visual and audio-only conditions [p=0.94]. Speech activity was statistically identical in all four conditions [mean=44.4%, sd=12.6 across all speakers].

4. Discussion

Noise affects speech production, and the 'noise' of a competing conversation tested here also led to Lombard

effects. However, talkers exhibited contrasting responses to overlapping speech from their interlocutor and that arising from the background. Simultaneous background speech had virtually no effect on the two prosodic parameters $F0$ and speech rate but did lead to an increase in speech output level. In contrast, prosodic parameters were heavily influenced by overlap with a conversational partner while no change in energy was seen. This outcome corroborates earlier findings [6] which highlighted the need to distinguish between Lombard and interactional influences on prosodic parameters.

Unexpectedly, speech output level did not increase with more than one active background talker, contrasting with our earlier study on competing conversations [6]. The seating configuration differed between the two studies: here, conversational partners faced each other around a circular table, while in [6] pairs of talkers sat next to each other. We hypothesise that in the current study, if speakers are to talk across the other pair, increasing speech level may not be an effective strategy to overcome noise, given that the other pair may be doing the same, resulting in a form of positive feedback detrimental to successful communication. Rather, speakers appear to be aware of the energetic masking potential of their speech on the other pair, and actively engage in minimising it. Shouting is not efficient nor cooperative in this scenario.

Globally, no clear effect of the visual modality was observed on the acoustic/prosodic parameters examined apart from a tendency for a slower speech rate in the audio-only condition. However, the degree of within-pair overlap tolerated decreased when visual contact was not permitted, and tended to reduce further with the background pair present. Overlap occurs naturally in fluid conversations, but it seems likely that overlap reduction is a strategic response from speaker-listener pairs to adverse conditions such as not being able to see the interlocutor, or having to separate a dialogue partner's speech from competing speech in the background.

It has been suggested that increases in speech level are necessary to maximise message reception when visual cues are absent [9]. However, we find here that talkers lowered their speech level in the specific times when overlapping with their interlocutor. Given the interactive nature of the masker where additional effort may be detrimental, speakers appear to have reacted by monitoring their output level in order to reduce masking for their interlocutor. Similarly, [14] observed that speakers took extra care to produce clearer prosodic contrasts when the visual modality was blocked.

Further, the fact that overlaps are proportions (computed with respect to overall speech activity) rules out the possibility that overlap reduction is the product of passive strategies such as speaking less in adverse conditions. Rather, speakers appear to retime their contributions so as to minimise interlocutor overlap.

While a previous study [15] found overlap reduction with the masker, here the overlap reduction is only found with the conversational partner. It seems likely that the high density of speech activity precluded attempts to exploit the infrequent epochs of silence, particularly given the imperative to maintain a conversation. It might be the case that speakers focused instead on their partner's speech instead of trying to reduce overlap with the masker. Indeed, some subjects reported that conversing in the presence of another conversation was *easier* when visual contact was blocked.

Acknowledgements. We thank Bill Wells and Emina Kurtic for useful discussions on conversation analysis. The conversational excerpt was formatted using a LaTeX package developed by Gareth Walker at the University of Sheffield. This work was supported by EU Future and Emerging Technology (FET–OPEN) Project LISTA (The Listening Talker).

5. References

[1] E. Lombard, "Le signe d'élévation de la voix," *Annales des maladies de l'oreille et du larynx*, vol. 37, pp. 101–119, 1911.

[2] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, no. 1, pp. 510–524, 1993.

[3] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1101–1109, 2001.

[4] S. Mattys, J. Brooks, and M. Cooke, "Recognizing speech under a processing load: Dissociating energetic from informational factors," *Cog. Psych.*, vol. 59, no. 3, pp. 203–243, 2009.

[5] 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, no. 7, pp. 936–941, 1962.

[6] V. Aubanel, M. Cooke, J. Villegas, and M. L. Garcia Lecumberri, "Conversing in the presence of a competing conversation: effects on speech production," in *Interspeech*, Florence, Italy, 2011, pp. 2833–2836.

[7] E. A. Boyle, A. H. Anderson, and A. Newlands, "The effects of visibility on dialogue and performance in a cooperative problem solving task," *Lang. Speech*, vol. 37, no. 1, pp. 1–20, 1994.

[8] Bull and Aylett, "An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue," in *ICSLP*, Sydney, 1998.

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

[10] E. Kurtic, "Overlapping talk and turn competition in multi-party conversations," Ph.D. dissertation, University of Sheffield, 2012.

[11] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Ver. 5.3.11, http://www.praat.org/, 2012.

[13] "BEEP dictionary," 1996. [Online]. Available: http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

[14] E. Cvejic, J. Kim, and C. Davis, "Effects of seeing the interlocutor on the production of prosodic contrasts," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. 1011–1014, 2012.

[15] M. Cooke and Y. Lu, "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2059–2069, 2010.