Conversing in the presence of a competing conversation: effects on speech production

Vincent Aubanel[1], *Martin Cooke*[1,2], *Julián Villegas*[1],
Maria Luisa Garcia Lecumberri[1]

[1]Language and Speech Laboratory, University of the Basque Country, Spain
[2]Ikerbasque, Basque Foundation for Science, 48011, Spain
{v.aubanel,j.villegas}@laslab.org, m.cooke@ikerbasque.org, garcia.lecumberri@ehu.es

Abstract

How does a background conversation affect a foreground conversation? In this scenario, and unlike traditional studies of noise-induced speech modification (Lombard speech), listeners have to cope with the additional challenge of competing speech material. In the current study, pairs of talkers engaged in natural dialogs in the absence or presence of another talker pair. Changes in speech level revealed only a small energetic masking effect of the background pair, but very large modifications in prosodic parameters ($F0$, speech rate) were observed during overlaps within conversations. The presence of the background pair led to increases in the numbers of dysfluencies, mistiming and interruptions, suggesting that interlocutors suffer from competing speech in ways which are not well-described by Lombard speech modifications. Longer inter-turn pauses seen in the background present condition may indicate that listeners monitor the other conversation to avoid temporally-competing speech material where possible.

Index Terms: Speech modifications, Lombard effect, Multiparty conversation analysis.

1. Introduction

Conversing in the presence of other conversations is an everyday occurrence. The background conversation can disrupt the foreground conversation by masking parts of it or by diverting attention. One question that arises is if and how speakers overcome these disruptions by engaging in strategies destined to favour a better reception of their message at the ears of their interlocutor. These strategies, if they exist, would provide useful insights into the mechanisms that underpin human speech communication. At an applied level, they could also lead to the development of speech technology applications that could simulate or even exceed human performance in successfully conveying a message in noisy environments.

Speech modifications occur in a number of situations e.g., when talking to children or foreigners [1], as a consequence of experimental manipulation [2], and in adverse conditions [3]. Research on speech production in noise has focused mainly on read speech or, more recently, on speech elicited through simple tasks [4, 5]. Very few studies have examined speech produced in the presence of competing speech, and while the informational masking effect of other speech on message reception has been studied (e.g., [6, 7, 8]), the possible role of informational disruption on speech production has been largely overlooked.

One early study which did investigate the effect of competing talkers on conversations [9] involved pairs of talkers communicating word lists. The presence of a second pair of talkers caused an increase in foreground speech level, a slight reduction in speech rate, and an increase in the number of communication errors. A recent task involving read speech [10] evaluated the effect of N competing talkers (N ranging from one to infinity— i.e., speech-shaped noise) on speech production. They found the extent of speech modifications increased with N, suggesting a correlation between the degree of the Lombard effect and energetic masking. In [11], speakers engaged in a communicative problem-solving task in quiet and noisy conditions, including a speech masker. They observed a reduction in overlap between foreground and background conversations relative to a quiet baseline, and suggested that speakers achieve this by monitoring the background speech and exploiting pauses to increase the chances of transmitting the message to their interlocutors.

Our aim in the current work is to investigate the effect of competing speech on a foreground conversation in the unrestricted setting of natural, spontaneous conversations. By avoiding the use of a specific task, speakers are under no pressure to engage in what may be artificial strategies to solve problems in a limited time. Instead, we would expect evidence for interactional resources to emerge which enable correct message reception. For example, listeners may signal more frequently (e.g. via back-channels) that they are following the conversation. At the same time, we are interested in identifying any negative effects of background conversations, such as on the precision of timing of turns as well as on listeners' ability to detect places where a turn-change is allowable [12].

Here we report on the design and collection of a competing conversation corpus and go on to examine both traditional Lombard effects on acoustic parameters as well as interactional aspects such as turn-taking.

2. Methods

2.1. Participants

Three pairs of female native speakers of Spanish took part in this study. Both members of two pairs knew each other well. All participants were students at the University of the Basque Country, and none reported any hearing problems.

2.2. Procedure

Each recording session started with a single pair conversing for five minutes, after which they were joined by the second pair, and both pairs then conversed simultaneously for ten minutes. For the final five minutes the first pair left the room while the second pair continued their conversation. Pairs took part in two

recording sessions. Participants sat at a table with a visual barrier placed in the middle as shown by Fig. 1. They were instructed to talk freely in pairs and not to engage conversation with the other pair. To elicit conversation they were given a list of neutral topics such as 'holidays' and 'shopping', but were allowed to talk about anything. In practice, conversation flowed freely for all recording sessions.

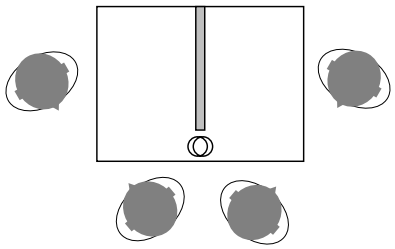


Figure 1: Experimental setting for the conversations: eye contact is possible within but not across pairs. Audio was also recorded with an omnidirectional microphone (◯).

Participants wore wired Sennheiser ME-3 head-mounted microphones connected to a MOTU 8Pre eight channel digital audio interface and were recorded on separate audio channels on a computer using AudioDesk software. Audio from an omnidirectional microphone (AKG 4500) was also recorded via the same mechanism.

2.3. Annotation

Speech material was annotated with MTRANS, a newly developed signal editor dedicated to multi-channel annotation [13]. First, each audio channel was manually segmented into silent/non-silent parts. These boundaries served as a basis for the semi-automatic calculation of *inter-turn intervals*, a cover term for *gaps*, acoustic silences occurring after a speaker finishes the turn and before the other speaker starts the next turn, and for *overlaps*, the portion of time at a speaker change when both speakers speak at the same time. The collection of gap and overlap durations lie on a continuum of inter-turn intervals (e.g., [14]).

Speech fragments were manually annotated into turn parts, following a turn-taking scheme adapted from [15] that aims at categorising any speech part into one of seven categories, describing the interactional structure of the dialogue. Apart from holds (*H*) and smooth changes (*S*), which together constitute the majority of turn maintenance and turn exchange, this scheme distinguishes between back-channels (*BC*) and three kinds of interruptive speech parts: interruptions (*I*), pause interruptions (*PI*) and butting-in (*BI*). Simultaneous starts (*SS*), when a speaker starts within 210 ms. after the other speaker and as a result it is not clear who owns the turn ([16]), are also coded.

Specific events relating to the interactional unfolding of the conversation were also annotated manually. These consist of elongations (*ELO*), e.g., vowels sustained for more than 300 ms., dysfluencies (*DYS*), denoting hesitations, stuttering or any interactionally-relevant disruption of speech, self-cutoffs (*CUT*) resulting mainly from interruption by the interlocutor, mistimings (*MIS*), where talkers enter the conversation at inappropriate points, and alignment phenomena (*ALI*), consisting of lexical, syntactic or intonational repetition patterns that are not necessary to the transfer of information for a successful interaction but rather display mutual understanding between speakers (e.g., [17]). Finally, a word level pseudo-orthographic transcription was also performed manually, and subsequently used for speech rate measurements.

Annotation levels are summarized in Table 1, alongside their frequency in the entire corpus.

Table 1: Labels, number of occurrences, and percentage of the different annotation levels.

Annotation level	label	N	%
inter-turn intervals		1093	
turn parts	S	658	29.5
	H	407	18.3
	BC	446	20.0
	SS	183	8.2
	BI	58	2.6
	I	168	7.5
	PI	307	13.8
	total	2227	100
events	DYS	464	
	MIS	174	
	CUT	170	
	ELO	512	
	ALI	116	
	total	1436	
words		18434	

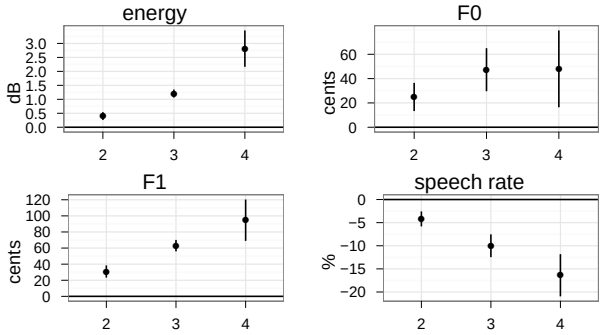
3. Results

3.1. Lombard effects

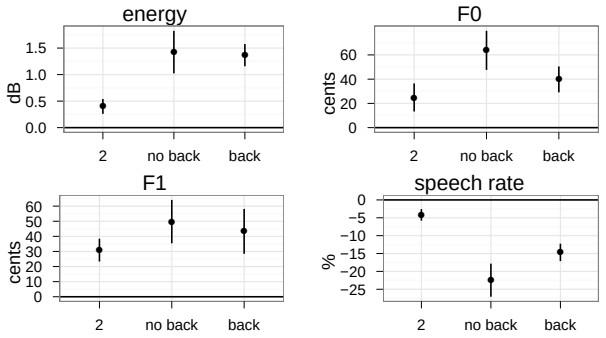
Using Praat [18] we computed energy, fundamental frequency ($F0$) and the frequency of the first formant (F1) on the speech parts of the corpus. As a proxy for speech rate we used the number of vowels per second derived from the word-level transcription. These parameters have been found to vary systematically with the presence of noise, and collectively represent the dominant elements of the Lombard effect [3, 19, 20].

Fig. 2 shows the results pooled over all combinations of speaker and recording session. To get an idea of the overall effect of an increase in background sound level, the upper panel plots the value of each parameter as a function of the number of active talkers, relative to the condition where only one talker was active. Increases in speech output level, $F0$ and F1, and decreases in speech rate are seen, as found in virtually all studies of Lombard speech. However, apart from speech rate, the changes are rather small compared to those observed in Lombard speech studies using non-speech noise backgrounds, suggesting that the energetic masking effect of the competing speech was relatively small. In contrast, the reduction in speech rate was substantially larger than seen in traditional Lombard studies.

The lower panel replots the 'any-two-talker' condition from the upper plot alongside values derived from overlapping speech from conversational pairs with the background absent or present. Relative to an arbitrary pair of talkers, stronger Lombard effects can be seen for interlocutors. In the background absent case, there is in fact no 'noise' inducing these modifications. Rather, the changes appear to be due to natural effects of dialogs on $F0$, speech rate, etc. Indeed, when the background pair was present (introducing 'noise'), the size of the modifications decreased.



(a)



(b)

Figure 2: Lombard effects for energy, $F0$, F1 and speech rate relatively to single speaker activity (a) for an increasing number of simultaneous talkers and (b) for pair overlaps without and with simultaneous background conversation. Error bars represent 95% confidence intervals over all speakers, which also applies for the subsequent analyses.

3.2. Interactional effects

3.2.1. Event counts

Fig. 3 shows the difference in frequency of events between the two different background conditions for each of the five event categories. A value greater than zero indicates a higher rate of occurrence when a background conversation is present, measured in events per minute. Speakers were found to produce significantly more dysfluencies (*DYS*) and mistimings (*MIS*) in the presence of a background conversation (*DYS*: $t(11) = 3.56$, $p < 0.01$; *MIS*: $t(5) = 3.13$, $p < 0.05$). No significant differences in the frequency of elongations (*ELO*), self-cutoffs (*CUT*) or alignments (*ALI*) were observed. A Generalised Linear Model analysis which included a factor representing the proximity of the speakers to the competing pair and presence/absence of background showed no evidence of significant influence of proximity in the experimental outcome.

3.2.2. Turn types

Fig. 4 presents the relative frequencies of turn types when the background speech was present relative to the background absent condition. On average, speakers interrupted more when a background conversation was present, both for interruptions when the other speaker was active (*I*: $t(11) = 2.23$, $p < 0.05$) and when the other speaker was pausing but without having indicated a preference to yield the turn (*PI*: $t(11) = 2.51$, $p < 0.05$). There was a tendency to see reduced back-channel

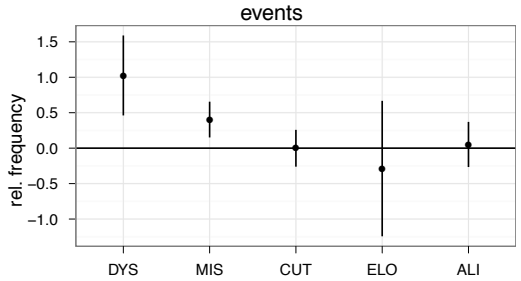


Figure 3: Event frequency in number of occurences per minute in background present relative to background absent.

(*BC*) usage in the background present condition, perhaps due to the availability of visual alternatives for signalling understanding.

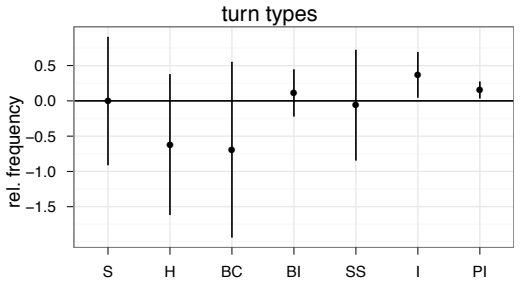


Figure 4: Turn type frequency in number of occurences per minute in background present relative to background absent.

3.2.3. Inter-turn intervals

Fig. 5 compares inter-turns interval durations grouped for all speakers for the two background conditions. Positive values represent pauses while negative values denote overlaps. Intervals are both more widely distributed and are positively shifted in the background present condition, as shown by tests for the variance and the mean: ($F(742, 349) = 1.888$, $p < 0.001$; $t(908.072) = 3.734$, $p < 0.001$). For each speaker pair in each session, standard deviations of interval durations were greater in the background condition. Inter-turn interval means of 97 ms and 234 ms for the background absent and present conditions respectively are in accordance with the finding that slight gaps are the most frequent between-speaker interval [12, 21]. The broadening at the tails of the distribution in the background present condition indicates a greater proportion of longer intervals, in particular for pauses.

4. Discussion and further work

That speakers modify their speech in challenging acoustic conditions has been known for a century [3]. The current study extends previous work on Lombard speech into natural conversations and takes the first steps towards quantifying the effect of a background conversation on a foreground conversation, allowing the analysis of both acoustic and interactional aspects.

Relative to traditional Lombard studies with unintelligible noise, a competing talker pair induces relatively small changes to parameters such as speech level, $F0$ and F1 frequency. This finding is compatible with models which show a clear relationship between the energetic masking potential of the noise back-

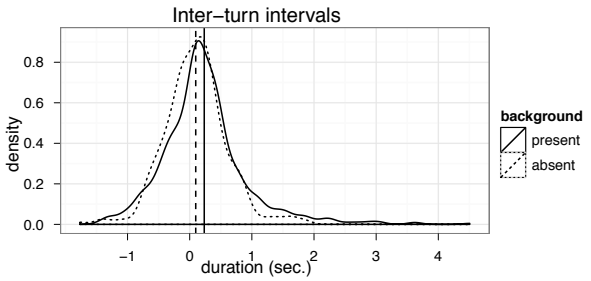


Figure 5: Densities of inter-turn intervals contrasting the two background conditions.

ground and the extent of speech modification [10], since competing speech is not a very effective energetic masker.

The scenario of the current study differs from the majority of Lombard studies in two ways: first, speakers were engaged in natural conversations; second, the masker material was intelligible competing speech. The first factor itself results in significant acoustic modifications which masquerade as Lombard effects as shown in the lower panel of figure 2 and as also found by [20, 22, 4]. We found very large effects on $F0$ and speech rate during the periods of overlapped speech even for speakers conversing naturally without the background pair present. Indeed, the presence of the background pair led to a partial suppression of these within-pair effects on $F0$ and rate, in spite of the increase in background noise energy relative to the background pair absent case. This is a clear indication that factors other than energetic masking are at play in determining the level of speech modifications in natural conversations.

A quantitative analysis demonstrated the adverse influence of the background conversation in several ways: an increase in dysfluencies and mistimings, greater numbers of interruptions, and the increased use of longer pauses (and to some extent longer periods of overlapping speech). The presence of longer pauses between turns echoes the findings of [11], who suggested that speakers monitor the background to determine when to start to speak. Recent models of dialogue [17, 23] argue that conversational partners employ predictive ('forward') models of their interlocutor's speech in order to engage in activities such as smooth turn-taking. In competing speech situations, it seems plausible that part of the same cognitive machinery could be deployed in deciding when to speak.

A further issue not addressed here concerns the role of visual information in maintaining conversations in the types of scenario explored here. For instance, here we were unable to determine whether the use of visual back-channels (e.g., nods) changes as a result of a competing conversation. We have recently extended the scenario to the collection of a larger-scale corpus of (English) conversations in conditions where partners are either able or unable to see their interlocutors.

In conclusion, the current study suggests that speaking in the presence of other talkers presents challenges for interlocutors, resulting in speech modifications which lie largely outside the framework of traditional acoustic Lombard speech effects. Deeper analyses at an interactional level are needed to determine what strategies are employed by speakers and listeners to maintain intelligibility in competing speech backgrounds.

Acknowledgements. We would like to thank Bill Wells, Emina Kurtic and Jan Gorisch for useful discussions about conversation analysis and overlaps. This work was supported by EU Future and Emerging Technology (FET–OPEN) Project LISTA (The Listening Talker).

5. References

[1] M. Uther, M. A. Knoll, and D. Burnham, "Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech," *Speech Communication*, vol. 49, no. 1, pp. 2–7, 2007.

[2] K. G. Munhall, E. N. MacDonald, S. K. Byrne, and I. S. Johnsrude, "Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 384–390, 2009.

[3] E. Lombard, "Le signe d'élévation de la voix," *Annales des maladies de l'oreille et du larynx*, vol. 37, no. 2, pp. 101–119, 1911.

[4] M. Garnier, "Communiquer en environnement bruyant: de l'adaptation jusqu'au forçage vocal," Ph.D. dissertation, U. Paris 6, Paris, 2007.

[5] Y. Lu, "Production and Perceptual Analysis of Lombard Effect," Ph.D. dissertation, U. of Sheffield, Sheffield, 2009.

[6] R. Carhart, T. Tillman, and E. Greetis, "Perceptual masking in multiple sound backgrounds," *J. Acoust. Soc. Am.*, vol. 45, no. 3, pp. 694–703, 1969.

[7] 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.

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

[9] J. C. Webster and R. G. 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.

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

[11] 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.

[12] 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.

[13] M. Cooke, J. Villegas, V. Aubanel, and M. A. Piccolino-Boniforti, "MTRANS: A MATLAB tool for multi-channel, multi-tier speech annotation," in *VLSP*, Pennsylvania, 2011.

[14] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *J. of Phonetics*, vol. 38, pp. 555–568, 2010.

[15] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, pp. 601–634, 2011.

[16] D. B. Fry, "Simple reaction-times to speech and non-speech stimuli," *Cortex*, no. 11, pp. 355–360, 1975.

[17] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. version 5.2.21, retrieved 29 march 2011 from http://www.praat.org/," 2011.

[19] 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.

[20] H. Lane and B. Tranel, "The Lombard Sign and the Role of Hearing in Speech," *J. of Speech and Hearing Research*, vol. 14, no. 4, pp. 677–709, 1971.

[21] J. Jaffe and S. Feldstein, *Rhythms of dialogue*. New York, NY, USA: Academic Press, 1970.

[22] J.-C. Junqua, "The Lombard effect: a reflex to better communicate with others in noise," in *ICASSP*, 1999.

[23] M. J. Pickering and S. Garrod, "Do people use language production to make predictions during comprehension?" *Trends in Cognitive Sciences*, vol. 11, no. 3, pp. 105–110, 2007.