Map Task Dialogs in Noise - a Paradigm for Examining Lombard Speech

Hansjörg Mixdorff, Ulrich Pech*, Chris Davis** and Jeesun Kim***

*Department of Computer Sciences and Media, Berlin University of Applied Sciences, Germany
mixdorff@tfh-berlin.de, UlrichPech@web.de

**MARCS Auditory Laboratories, University of Western Sydney, Australia
{chris.davis;j.kim}@uws.edu.au

ABSTRACT

This paper presents a paradigm for comparing auditory-visual map task dialogs produced in silence and in noise, also known as Lombard speech. A previously developed temporal filtering algorithm which removes the ambient noise from recordings of Lombard speech by locating and subtracting a recording of the noise performed in the same environment was modified to accommodate longer recordings. The filtering algorithm yields overall noise attenuation between 15 and 35 dB without distorting the speech signal like spectral filtering approaches. On a small production dataset of two levels of vehicle and babble noise we examined the effect on fundamental frequency and intensity contours. We found that Lombard characteristics of speech, that is, an increase in mean *F0* as well as intensity, are stronger for babble than for vehicle noise. There are indications that talkers become habituated to the noisy environment when they are exposed to it for the duration of a dialog. We did not find any consistency regarding the speed of completion of the map task, although participants appeared to solve the task more leisurely in silence than in noise. By performing eye-tracking on one of the talkers' data we found that the frequency of gaze was more than double in babble noise than in silence.

1. INTRODUCTION

It is commonly known that humans in noisy environments adapt their manner of speaking. This adaptation not only affects the loudness of speech, but also fundamental frequency, speech rate and spectral characteristics (for a summary see, for instance, [3]).

The current study examines auditory-visual dialogs and the way they are influenced by ambient noise of various kinds and levels. We aim to investigate whether apart from the typical acoustic modifications observed in Lombard speech also other aspects of the auditory-visual communication change, that is, for instance, talkers' dialog strategies.

To this effect we employ a filtering approach developed by the first author [4] to remove the ambient noise from recordings of Lombard speech. In [4] we had examined the feasibility of the filtering approach on a small corpus of phonetically balanced isolated sentences of Finnish and German. We found that the filtered speech data were clean enough to reliably perform acoustic analyses, such as *F0* and formant frequency extraction.

In order to study the effect of noisy environments on a problem-solving task we adopted the well-known map task paradigm which has been studied and documented for a large number of languages [1]. The map task can be performed with or without eye-contact. In our case the participants were sitting at the same table and had eye-contact. Both maps were presented on notebook computers.

2. METHODS AND SPEECH MATERIAL

We first give a short description of the filtering algorithm. We assume that the recording set-up is highly linear, time-invariant and low in noise, and therefore its transfer function remains fairly constant. While the talkers are already in the room a sample of the contaminating noise is played back and recorded through the microphones which are subsequently used for the speech recording. In the case of isolated sentences [4], noise duration of four seconds was sufficient.

As experience shows, typical map tasks can take between 2 and 15 minutes. Since it was unfeasible to record noise samples of two minutes with the talkers remaining silent, we decided to use noise intervals of 10 seconds and loop these for the required duration. Therefore, every trial started with 10 seconds of noise playback

followed by a silence of 5 seconds. Then the noise playback resumed and shortly after, the participants were signaled to start the map task. After a recording is completed, the initial noise-only recording is subtracted from the contaminated speech samples.

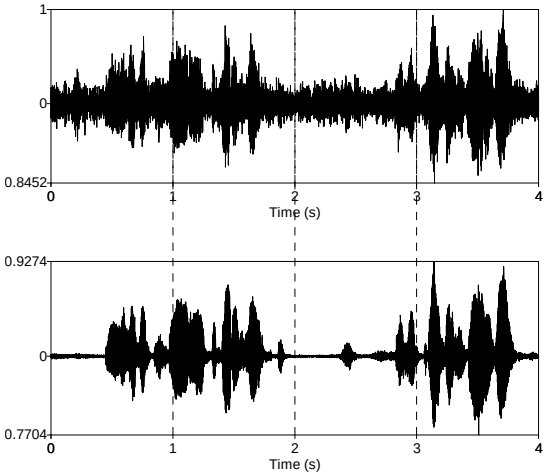

Figure 1: Noise-contaminated speech sample before and after filtering, speaker M, condition *babble*.

It should be noted that the attenuation level may vary by up to 6 dB as the participants move their heads or hands, for instance, during the course of the map task and hence modify the transfer function of the setting. Figure 1 shows a typical example from the data before and after filtering.

The two participants sat in a sound-proof room facing each other. Each of them was video-recorded on DV cassette (576 x 720 pixels, 25 frames/second) as well as audio-recorded using a PC at 16 kHz/16 bit. The monophonic ambient noise was played back in synchrony with the recording over two loudspeakers. Auxiliary microphones were used to record the audio in the room onto the video tapes in order to later on synchronize the video with the audio stream. The maps were displayed on notebook computers. The activity of the follower was monitored using the desktop capture software *1stScreenRecorder*[5].

We chose the following conditions for examining possible changes in the participants' acoustic speech characteristics as well as - potentially - their strategies when solving the map task (notations henceforth used in italics):

1) Silence, *silence*
2) Vehicle noise, SPL=84 dB(A), *vehicle*
3) Vehicle noise, 78 dB (A), *vehicle -6dB*
4) Babble noise, 84 dB (A), *babble*
5) Babble, 78 dB (A), *babble -6dB*
6) Whisper in silence, *whisper*

The maximum noise levels were set to a value at which participants noticeably raised their voices but did not need to actually yell at each other.

Considering that we had six conditions to compare, the nature of the map task posed certain problems. Two constraints needed to be addressed: On the one hand, due to potential learning effects, a set of maps that has already been worked on by the participants should not be re-used. On the other hand, for the sake of comparability, the landmarks as well as the structure of the task should not be changed between trials. Therefore we developed mirrored and/or rotated versions of one and the same map. We hoped that the modified spatial relationships would compensate the learning effects.

In the first trials, from which we report results in this article, we used sets of maps, in which the giver and the follower had slightly different versions. Therefore the participants resolved the discrepancies during the first trial and were relatively faster in all the following ones. In future stages of the study we plan to use exactly the same landmarks in both maps and only change the map orientation. Before the actual experiment we ran a warm-up trial in which a completely different set of maps was presented in order to familiarize the participants with the task. For our experiments we employed maps which were developed for German by K. Claßen [2] and translated them to English.

For the purpose of the current article which is chiefly concerned with the development of this paradigm we present results from two sets of map tasks, each from a different pair of talkers who are native speakers of Australian English and members of staff at MARCS Auditory Labs. We concentrate on the speech data recorded from the two givers, henceforth talker *F* (female) and talker *M* (male).

After filtering the speech samples, fundamental frequency contours were extracted at a time-step of 10 ms using *Praat* [6]. Since the talkers were sitting in the same room the crosstalk level between them was relatively high (typical attenuation of 15 dB). This occasionally posed problems for the automatic *F0* extraction. The data were therefore inspected and if necessary

corrected. Intensity values were calculated by first segmenting the data into speech and non-speech segments and then calculating *Praat Intensity* contours for the speech segments. The DV data which were filmed in portrait mode to yield a better resolution of the participants' faces was captured onto a PC, de-interlaced and rotated using *VirtualDub* [7].

Subsequently, the filtered audio streams, the participants' videos as well as the screen capture video of the follower's map were time-aligned in Adobe Premiere 6.5 and output in a format of 900 x 360 pixels at 25 frames per second for later analysis.

3. RESULTS OF ANALYSIS

Table 1 shows the durations of the six different map tasks. It must be noted that due to the fact that the first task involved the disambiguation of landmarks it took relatively longer to complete.

Given this potential "warm-up" effect, the data presented below may result from the combined effect of acoustic environment and habituation to a specific map. Taking this into account, it comes as a surprise that the tasks worked on in silence took relatively longer compared to those in noise. The same applies to the whispered map tasks. It seems that participants in the silent environment were more at ease and negotiated the task more leisurely. However, much more data with varying presentation orders are needed in order to determine the contributions of noise type and level on performance speed.

Table 1: Order of presentation (in parentheses) and durations of map tasks.

condition	talker F	talker M
vehicle	(1) 316.3 s	(2) 84.2 s
vehicle -6dB	(4) 204.6 s	(4) 83.6 s
babble	(5) 170.4 s	(1) 107.7 s
babble -6 dB	(6) 170.9 s	(6) 96.9 s
silence	(3) 243.6 s	(5) 110.4 s
whisper	(2) 343.2 s	(3) 100.8 s

Table 2 shows means and standard deviations of *F0* for all conditions. The Lombard effect is clearly marked by rises in the mean *F0* for both talkers, and also by a rise in the standard deviation for the female. The effect is stronger for babble than for vehicle noise. In Table 3, *F0* minima and maxima as well as 10 and 90% quantiles of *F0* are given.

Table 2: Mean/standard deviation of F0 in Hz.

condition	talker F	talker M
vehicle	283.0 / 57.9	208.5 / 34.8
vehicle -6dB	249.3 / 45.4	196.5 / 38.5
babble	320.4 / 60.9	224.0 / 30.3
babble -6 dB	294.2 / 57.2	208.6/ 35.8
silence	205.0 / 37.5	160.8 / 35.7

Table 3: Minima/10%quantile/90%quantile/maxima of *F0* in Hz.

condition	talker F	talker M
vehicle	149/227/352/599	142/180/247/383
vehicle -6dB	151/209/293/593	107/166/231/385
babble	167/259/389/744	153/198/257/397
babble -6 dB	155/243/355/599	114/182/250/376
silence	149/173/245/502	100/129/200/389

The data suggest different strategies as the ambient noise becomes stronger: Whereas female talker F raises the upper limit of her *F0* range, the maximum value used by male speaker M is fairly constant even when comparing conditions *silence* and *babble noise*. Instead, he shifts his lower *F0* limit. The fact that his *F0* range is compressed as the noise level rises is also reflected by a decrease in the mean absolute slope of his *F0* contours as displayed in Table 4. The relationship is reversed for talker F.

As we were interested in determining whether some sort of habituation occurred as the task proceeded we examined if the local mean value of *F0* decreased with time. To this end we calculated *F0* means for frames of 20 seconds at a time-step of one second. We chose the large frame size in order to smooth out possible discourse related variations. Although in most "noisy" cases the first frame had the highest *F0* mean, this could as well be attributed to the opening phase of the task. We correlated the mean frame *F0* with the point in time where it occurred to arrive at the data displayed in Table 5.

Table 4: Mean absolute *F0* slopes in semitones.

condition	talker F	talker M
vehicle	26.3	17.3
vehicle -6dB	18.0	20.6
babble	26.3	13.4
babble -6 dB	22.8	19.8
silence	16.7	32.3

Table 5: Correlation between time into trial and mean *F0* calculated for frames of 20 sec at a time-step of 1 sec.

condition	talker F	talker M
vehicle	-.208 (h.s.)	-.727 (h.s.)
vehicle -6dB	.465 (h.s.)	.354 (h.s.)
babble	-.270 (h.s.)	-.639 (h.s.)
babble -6 dB	.313 (h.s.)	-.274 (h.s.)
silence	.292 (h.s.)	-.084 (n.s.)

Although we found highly significant negative correlations for the majority of the noisy conditions, it is hard to explain why some showed highly significant positive correlations. One possibility is that habituation only occurred at the highest noise levels since there is more evidence of a negative correlation if only the cases of maximum vehicle or babble noise are considered.

Table 6 gives intensity means and standard deviations in dB. As expected, the talkers speak louder in a noisy environment, as well as when the noise level is raised. As in the *F0* data, the babble noise causes a stronger Lombard effect than the vehicle noise. For the female talker we also observe a reduction in the standard deviation.

The video data of talker M were analyzed for cases *silence* and *babble* in order to examine whether the noisy environment increased the frequency of gazes at the follower. To this end, an image tracking software was employed that traced the movements of talker M's pupil relative to a rigid point (tip of the nose). Using one frame where talker M fixates his map and another one where he gazes at his partner as references, we established a threshold above which a frame was counted as a "gaze frame". In the silent condition only 6.35% of video frames were counted as potential gazes, against 14.11% in the babble condition. The same was attempted for the follower but yielded unsatisfactory results because she was wearing glasses.

4. DISCUSSION AND CONCLUSIONS

This article presented a pilot study examining auditory-visual map tasks in noise. To investigate this we developed a set of recording and signal treatment procedures that do not interfere with a talker's auditory feedback. Although the data examined so far are very limited, it can be stated that the noisy environments induce the typical characteristics of Lombard speech, that is, raised *F0* and intensity, with babble noise being the stronger masking signal. Whereas the female raised her *F0* ceiling, the male raised his *F0* floor. In the cases of maximum vehicle or babble noise we found a negative correlation between time into task and frame mean *F0* that suggests a certain habituation of the participants over time. If we exclude the untypical first trials, which served as a kind of warm-up, participants are faster at solving the map tasks when they are exposed to ambient noise than in silence. This suggests that the noisy environment forces talkers to be more concise.
They took the longest in the condition *whisper*. Gaze analysis on the video data suggests a higher frequency of gazes in noise.

Table 6: Intensity mean and standard deviation in dB.

condition	talker F	talker M
vehicle	66.88/7.10	69.86/7.02
vehicle -6dB	63.08/7.54	67.88/7.68
babble	70.72/6.48	73.06/8.29
babble -6 dB	69.63/8.05	69.92/8.93
silence	62.72/10.72	54.44/8.73
whisper	53.72/9.21	51.43/7.83

In future works we will reexamine the results yielded so far on a larger group of subjects and also have a look at turn-taking strategies as well as visual interaction.

5. ACKNOWLEDGEMENTS

This work was supported by joint DFG/ARC cooperation grant 447 AUS-111/1/06 to the first author, as well as internal funding by MARCS Auditory Laboratories. Special thanks go to MARCS Lab Manager Colin Schoknecht for technical support as well as to MARCS research assistants who took part in this pilot study.

6. REFERENCES

[1] Brown, G. Anderson, A.H., Yule, G., & Shillcock, R. 1984. *Teaching Talk*. Cambridge, England: Cambridge University Press
[2] Claßen, K., 2000. Map Task – Eine Version für das Deutsche. AIMS, vol. 6 (4), pp. 65-83.
[3] Junqua, J.-P. 1996. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication* 20, 13-22.
[4] Mixdorff, H., Grauwinkel, K. Vainio, M. (2006): Time-domain Noise Subtraction on Lombard Speech. Proceedings of *Speech Prosody 2006*, Dresden, Germany.
[5] www.screenrecorder.us
[6] www.praat.org
[7] www.virtualdub.org