Content Realization in Noisy Environments

Magnus Sahlgren & Fredrik Olsson
Swedish Institute of Computer Science, SICS
{mange|fredriko}@sics.se

January 13, 2005

Abstract

It is well known that a noisy environment affects the prosodic and acoustic-phonetic properties of speech in the sense that speakers raise their voice, speak slower and increase their vocal pitch. This way of monitoring and compensating the vocal output is known as the Lombard reflex. The aim of the present study is to find out if and how the Lombard reflex extends beyond the acoustic-phonetic properties of speech to influence the way the communicative content is realized. Does a noisy environment affect the way people linguistically and pragmatically realize their utterances? We conducted a small user study in which two test subjects communicated over a noisy channel regarding an assembly task. The dialogues were recorded and analyzed. The results suggest that, at the linguistic level, speakers in noisy environments move towards structural completeness, while at the pragmatical level, speakers drift towards more active communicative participation.

1 Introduction

The effects of noise on speech production are well-known. People will inevitably raise their voice in the presence of noise — we have all been in situations where increased background noise makes us talk (much) louder than we usually do: bars, subways, airplanes, concerts, etc. are all notorious voice-raisers. The reflex to speak louder in the presence of noise is known as the *Lombard effect*, and was first described in 1911 by Etienne Lombard [4]. A fair amount of research has been devoted to the phonetic aspects of this phenomenon. For a good background on the early research, see the works of Lane and colleagues in the 1970s [3, 2]. For more recent surveys, see Summers et al., 1988 [5] and Junqua, 1993 [1].

The fact that speakers modify the amplitude of their voice in noisy environments suggests that speakers monitor their vocal output very carefully when speaking. It is clear from research on the Lombard reflex that this applies to the acoustic-phonetic aspects of speech. The question is whether it also applies to the *content* of speech? Does a noisy environment affect the way people linguistically or pragmatically realize their utterances? It seems fair to believe that people not only become articulatory more legible in noisy environments — i.e. speaking louder and slower — but that they also become linguistically and/or pragmatically more concise. Exactly what this means is debatable, but one might conjecture that it may involve simplification — morphologically, grammatically, syntactically, semantically and pragmatically — in the presence of noise.

In this paper, we wish to experimentally investigate whether people realize the content of their utterances differently when speaking in noisy environments. If this turns out to be true, we want to identify the linguistic and pragmatic features that are affected by Lombard speech. In order to accomplish this, we perform an assembly-type user study, where two test subjects communicate over a noisy channel regarding an assembly task. We record the dialogues under two noise conditions (no noise/lots of noise), and we analyze the dialogues with respect to content realization. Our experiments indicate that the Lombard reflex not only applies to phonetic features of speech, but that it also applies to the realization of content. We identify four pragmatic features, and three linguistic (structural) features that are clearly affected by Lombard speech. Our conclusion is that the Lombard reflex affects pragmatic features more than the linguistic ones.

Before we turn to the subject of content realization in noisy environments, we briefly review the main prosodic and acoustic-phonetic properties of Lombard speech.

2 Phonetic Realization in Noisy Environments

As noted in the introductory section, the Lombard effect has enjoyed considerable attention since it was first described. Below we list some of the most salient phonetic properties of Lombard speech that have been reported in the literature [3, 2, 5, 1]:

- Reduction of speaking rate.
- Increase in duration and intensity of utterances.
- Increase in vocal pitch.
- Vowels are lengthened, and consonants are shortened.

A somewhat surprising effect of the above characteristics of Lombard speech is that speakers tend to be more intelligible when subject to noise than when speaking in a quiet environment. Perception experiments have shown that test subjects understand Lombard speech more readily than "normal" speech [5, 1]. However, the level of intelligibility is not linear with the vocal effort of speakers, since shouting is (unsurprisingly) less intelligible than Lombard speech.

Apart from the above listed characteristics that seem more or less undisputed, there have been reports of other phonetic effects of Lombard speech that are not unanimously accepted. A severe problem with research in this area is that different studies more often than not come up with different, occasionally conflicting, analyses, and that the experiments are difficult, if not impossible to replicate. As an example of a study, Summers et al. [5] used 2 test subjects whose task was to pronounce 15 words from the Air Force speech recognition vocabulary under four different noise conditions: quiet, 80dB masking noise, 90dB masking noise, and 100dB masking noise. The authors report the standard Lombard reflexes including increase in amplitude and increase in duration. In addition to these properties, Summers et al. report the following effects:

- Increase in fundamental frequency.
- Decrease in *spectral tilt*, which is a measure of the relative energy at high and low frequencies. Thus, the relative distribution of spectral energy changes under noisy conditions, so that the high frequency end of the spectrum has more energy than the low end of the spectrum when noise is present. Summers et al. argue that this is not only an effect of increasing F0.
- Increase in vowel formant frequencies. In their study, Summers et al. note that F1 appears to be more affected than F2, and that there appears to be some correlation between F0 and F1.
- Flattening of the vowel spectrum. The vowel spectrum appears to be more flat in the presence of noise than in quiet conditions. The spectral flattening correlates with the spectral tilt.

These findings are not invariably supported by other studies of the Lombard reflex. For example, Junqua [1] observe that, while most speakers *do* exhibit the same general pattern as that observed by Summers et al., Lombard effects tend to be highly speaker (and also gender) dependent. Not only does the reflex vary across speakers; it also varies *within* individual speakers from one occasion to another. This means that, although it might be possible to identify general Lombard tendencies — such as the ones we have already mentioned, one should exercise great caution when trying to excavate more specific acoustic-phonetic changes. This caveat applies particularly to experimental studies that generalize from samples of speakers.[1]

3 Content Realization in Noisy Environments

As we have seen in the previous section, the phonetic aspects of the Lombard reflex have been thoroughly investigated. However, the linguistic/pragmatic aspects have not. If the assumption is correct that speakers monitor their vocal ouput carefully enough to consciously modify the phonetic realization, it is fair to believe that speakers also consciously modify the content realization. Whether this is so is the subject of the following study.

3.1 Assembly-Task Experiment

The investigation took place as a set of experiments concerning an assembly-task involving two subjects who communicated using speech only. The task the subjects were set out to solve was that of collaborating in building abstract objects of Lego, which were similar only in terms of number of Lego pieces involved. Participant A was given instructions in the form of photos for two different objects (see Figure 1)[2], while participant B was given more than enough Lego to build the corresponding objects.

[1] Needless to say, we are aware of its applicability to the present experiment.

[2] The abstract Lego objects reflect the hermeneutic foundation of the transcendental phenomenological reduction of *Das ding an sich*, and we have named the objects "Enigma", and "The Desolate One".

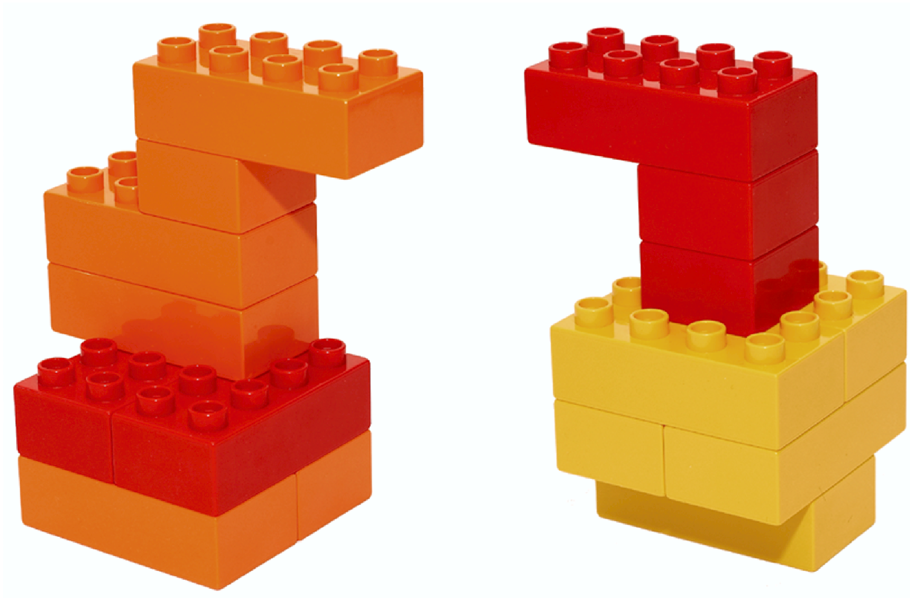

Figure 1: The instructions for the objects to be built were handed to the subjects in the form of photos.

In total, the investigation involved four subjects, all male, participating in two sets of experiments, each consisting of two tasks — i.e. two pairs of test subjects solved two assembly tasks each, one under noise, and one with no noise. The experiments were conducted using two computers, one for each participant, located in different rooms. The subjects communicated using Skype[3], a free voice-over-IP telephony software, and Philips head-sets[4]. On both computers, the SoundMasker software[5] was used to generate continuous noise according to the settings illustrated in Figure 2. WaveSurfer[6] was used on each computer to record the conversation between the subjects, and Audacity[7] was used during the analysis and alignment of the data (see Figure 3).[8]

The experiments were conducted under a not-so-very-rigorous four minute time constraint, or to the point of task completion (whichever occurred first).

[3]Skype is available at `http://www.skype.com`.

[4]Model SBC HM385.

[5]SoundMasker is available at `http://www.relaxingsoftware.com/`.

[6]WaveSurfer version 1.7.4 is available at `http://www.speech.kth.se/wavesurfer/`.

[7]Audacity version 1.2.3 is available at `http://audacity.sourceforge.net`.

[8]We initially envisaged a quite different experiment setup. Our first attempt was to use a stand-alone mixer, and to connect both channels (speaker A and speaker B) through the mixer, and to record the mixed dialogue. However, such a simple setup proved to be incredibly difficult to realize. We spent roughly three days trying, during which we visited the DIY (Do-It-Yourself) store Clas Ohlson several times a day, bought 16 or so connector plugs (stereo *and* mono — at one point, the shop assistant told us “I have never seen such a plug”), a number of cords, and the occasional adapter. We finally gave up. If nothing else, we had to, in order to maintain our sanity. Our second thought was to use a three-way Skype conference call, where the third party would be recording the dialogue. This also failed miserably — the host computer continuously crashed, for reasons we have yet to understand. Lesson learned: setting up a working environment for conducting a user study takes far more time than one might first think.

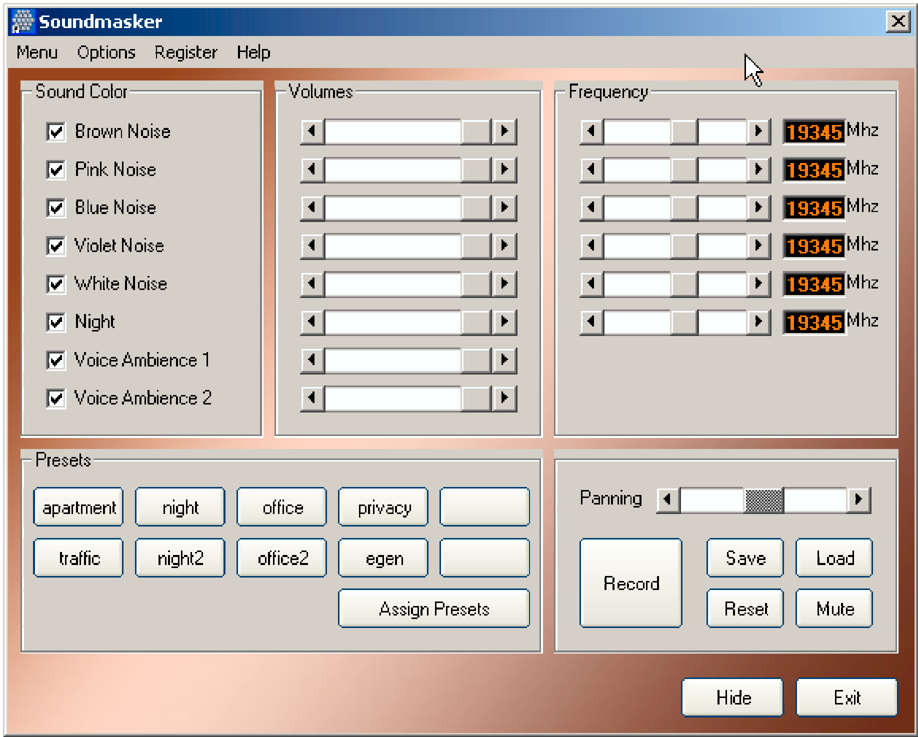

Figure 2: SoundMasker settings used to generate the noise for the experimenst.

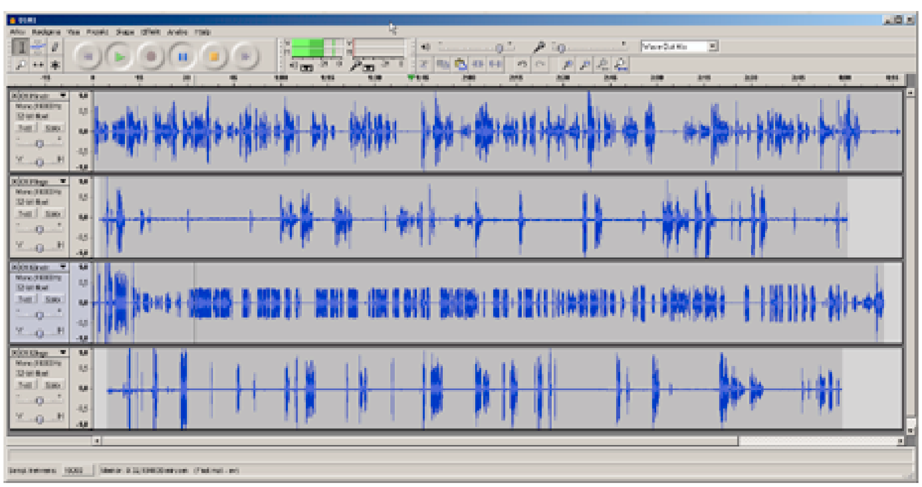

Figure 3: An example of Audacity as used for analyzing the conversation between two subjects under noiseless conditions (top two spectrograms) and noisy conditions (bottom two spectrograms). The first spectrogram for each pair is that of the instructor.

Since each pair of test subjects conducted two tasks, there was a high probability that the second assembly task would be experienced by the test subjects as being easier than the first one simply because they have gotten familiar with the task setting. This of course will affect the communicative realization in various ways. Since we do not want our results to be overly influenced by this dependence factor, we changed the order in which the test pairs were subjected to noise. The first pair started with a quiet setting, while the second pair started with a noisy setting.

All four conducted assembly tasks reached completion, and three of them were "correct" (i.e. the resulting Lego objects corresponded to the intended design).

3.2 Results and Analysis

Our main interest in this paper is to identify content features that are affected by the Lombard reflex. In order to do so, we performed auditory analysis of the dialogues with the intention of identifying linguistic/pragmatic features that changed in some significant manner in the noisy setting. It proved to be extremely difficult to precisely quantify the differences between feature realizations in the respective dialogues. In order to arrive at more precise numbers, it would have been necessary to first transcribe the dialogues, and then to comparatively analyze the transcriptions. However, since transcription is a very time-consuming business, we instead listened very carefully to the dialogues while taking notes and counting the occurrences of the linguistic and pragmatic features outlined in the tables 1 and 2 below. As noted at the end of section 2, we are fully aware of the dangers of generalizing based on few samples. We choose to view the following analysis as a first tentative step towards the identification of content-based Lombard features.

3.2.1 Structural Lombard Effects

As a general tendency, sentences uttered in the noisy setting appeared to be more complete, and to possibly be structurally simpler. In the quiet setting, the sentences tended to be more incomplete, with lots of subordinate clauses, restarts, and hestitations. In the noisy setting, on the contrary, sentences were predominantly more complete, with fewer subordinate clauses, and distinctively fewer restarts, and hestitations. However, the few pauses and hesitations that occurred in the noisy setting were often strongly voiced as if to signal that the channel was still open.

3.2.2 Pragmatic Lombard Effects

The Lego builders were more active in the dialogue in the noisy setting, and provided constructive feedback to the instructors. There were several different kinds of feedback in the dialogues. *Acknowledging* feedback — such as "ja", "jag förstår", "ok" — was generally increased when noise was added. This kind of feedback also tended to be more distinct in the noisy setting, and typically consisted of only one word ("ja", "ok"), or a short phrase "jag förstår"). In the presence of noise, the Lego builders also used more of what we call *align* feedback, which is basically control questions, such as "den gula?", which are used to align the dialogues, and to *ground* the task. The last, and possibly most significant, feedback type is what we call *proceed* feedback, which signals to the instructor to move on in the task. The Lego builders would often summarize their actions up to the point of the feedback, and then make the instructor understand that he should continue with a new subtask.

Another interesting pragmatic effect of Lombard speech was that the dialogues became more instructional, in the sense that the instructors were very clearly the dominating partner throughout the dialogues in the noisy setting. The Lego builders used their turns more for confirmation and feedback, whereas the instructors used their turns predominantly for giving concrete instructions. It was clearly the instructors' conceptualizations of the task that were the norm in the noisy setting. The quiet setting was definitely more cooperative in the sense that the test subjects negotiated more regarding the task conceptualization. Here, the Lego builders would occasionally enforce their understanding of the intended Lego object design.

3.2.3 Comparative Analysis

Having established a tentative inventory for content-based Lombard features, we analyzed the dialogues with respect to the apparent differences in the quiet and noisy settings. The results form the initial quantitative analysis illustrates the difference in feature occurrences between the quiet and noisy settings. Table 1 and Table 2 present the differences between the noisy and noiseless settings in terms of absolute numbers of realization of grammatical and pragmatic Lombard features.

Table 1: The difference between the noisy and noiseless settings in terms of absolute numbers between grammatical Lombard features. A and C are the instructors.

Speaker	Complete sentences	Incomplete sentences	Hestitations	Restarts
A	+15	-2	-1	0
B	+14	-6	-1	-3
C	+14	+7	+3	+7
D	+17	-2	-1	-1

As Table 1 indicates, there appears to be a shift towards more complete sentences in Lombard speech (for instance, in the noisy setting the number of complete sentences uttered by speaker A increased by 15 as compared to the noiseless setting). For all speakers in our experiment, the number of complete sentences were increased in the noisy setting, and for all speakers except one, the number of incomplete sentences were decreased. Also, there appeared to be a tendency to hesitate, pause and restart sentences more in the quiet setting. The fact that speakers seem to be in some sense *more grammatical* in Lombard speech supports the idea that Lombard speech tends to be more intelligible.

Regarding the pragmatic Lombard features, Table 2 indicates that there seems to be a general increase in feedback in the noisy setting. All speakers were more active in the noisy setting, and the intensity of turn taking was clearly higher. An interesting thing to note is that the instructors (A and C) were

Table 2: The difference between the noisy and noiseless settings in terms of absolute numbers between pragmatic Lombard features.

Speaker	Acknowledge	Align	Proceed	Initiative
A	-1	+2	-1	+12
B	+5	-7	+6	0
C	+4	+13	0	+14
D	+8	+3	+2	0

significantly more dominating in the noisy setting, and had the communicative initiative throughout the dialogues. Another interesting aspect of the analysis is that the Lego builders (B and D) used significantly more constructive feedback in the noisy setting. The "Proceed" feature in Table 2 is feedback that is intended to signal subtask completion, and to move the dialogue on. There were no "proceed" utterances in the quiet setting, but a number of them for both Lego builders in the noisy one.

As a general observation regarding the difference between normal and Lombard speech, there appeared to be more overlapping speech in the noisy setting. We believe that this is not only an artefact of the poor auditory conditions, but that it also depends on slower cognitive processing during Lombard speech; it is simply more difficult to process the input when there is a considerable amount of noise in the signal. The resulting phenomenon is reminiscent of the lag that can be experienced when using long-distance telephone lines.

4 Conclusions

Our main goal in this study has been to identify content-based Lombard reflexes. Despite the admittedly somewhat poor data, we managed to identify two main types of content-based Lombard features: grammatical and pragmatical, where the former basically means a drift towards structural completeness, and the latter means a drift towards more active communicative participation. We also presented a brief and tentative comparative study between normal and Lombard speech that demonstrated the occurrence of the mentioned features.

As for future work, it would be interesting to transcribe the dialogues, and to compare frequencies and word distributions in the respective noise conditions. It would also be interesting to perform a similar, but larger study, based on the findings reported here and applied to a different kind of disturbance; that of silence, clipping and noise characteristic for the cell phone network. We submit this idea as a suggestion and an appeal for future research into the content-based factors of Lombard speech.

References

[1] J. Junqua. The lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America*, 93:510–524, 1993.

[2] H. Lane and B. Tranel. The lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, 14:677–709, 1971.

[3] H.L. Lane, B. Tranel, and C. Sisson. Regulation of voice communication by sensory dynamics. *Journal of the Acoustical Society of America*, 47:618–624, 1970.

[4] E. Lombard. Le signe de l'elevation de la voix. *Annals Maladies Oreille, Larynx, Nez, Pharynx*, 37:101–119, 1911.

[5] W.V. Van Summers, D.B. Pisoni, R.H. Bernacki, R.I. Pedlow, and M.A. Stokes. Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, 84:917–928, 1988.