The Lombard Effect in Spontaneous Dialog Speech

Laura Folk, Florian Schiel

Bavarian Archive for Speech Signals, Institute for Phonetics and Speech Processing,
Ludwig-Maximilians-Universität, München, Germany

laure|schiel@phonetik.uni-muenchen.de

Abstract

The Lombard effect – environmental noise affects speech production – has already been studied extensively for read lab speech. In this study spontaneous dialog speech produced by 24 German speakers has been recorded under noisy conditions and analysed for the Lombard effect. A sophisticated experimental setup using behind-the-ear hearing aid equipment allows us to insert real car noise into the perceived audio stream of speakers while maintaining the normal auditory feedback loop. We found that the main Lombard effects – rising fundamental frequency and intensity – can be confirmed for dialog speech. Speaking rate did not slow down although reported earlier for read speech. We also found that certain rhythmicity features regarding the dynamic of the RMS energy contour change significantly under Lombard conditions but only for the female speakers.

Index Terms: Lombard effect, dialog speech, spontaneous speech, fundamental frequency, energy, rhythmicity features

1. Introduction

In many real-life situations speakers have to communicate under impaired conditions, i.e. noisy environments of any kind. In such situations, most speakers raise their pitch, increase the intensity of their speech signal and possibly reduce their speaking rate. This reflexive reaction became well known as the Lombard effect, first described by Étienne Lombard in 1911 ([1]). Since then many investigations have confirmed his observations; most have analysed laboratory speech masked with white noise for the interfering conditions (e.g. [2], [3], [4], [5]). All studies have shown that speakers gradually rise their fundamental frequency when the signal-to-noise-ratio decreases, and at the same time increase the intensity of their speech signal[1]. The latter roughly follows a linear function with a slope of about 0.5 (in decibel units), which means that speakers compensate an increase of noise intensity by increasing their own intensity by half the perceived increase ([2]).

Furthermore, some studies found that speaking rate is decelerated under noisy conditions which is an obvious strategy for remaining intelligible (e.g. [3]).

These findings are consistent with Lindblom's H-H theory ([6]) which postulates a continuum between "clear", i.e. hyper-articulated and relatively slow speech versus "reduced", i.e. hypo-articulated speech which is relatively fast, but economic. Lindblom's theory is founded on the notion that motor control of speech production is a trade-off between maximal communication vs. minimal effort and is therefore extremely variable according to different contexts and situations. The Lombard effect is therefore simply a reflex to shift the motor control of speech production in the direction of hyper speech to ensure an undisturbed communication. This hypothesis is supported by the fact that under noisy conditions listeners perceive speech which itself was produced under noisy conditions better than speech which was produced in quiet ambience ([7]).

Quantitative knowledge about the influence of the Lombard effect on F0 plays an important role in forensic investigations, since fundamental frequency is a key feature for speaker recognition.

Existing research has been based on read speech and unnatural white noise. To date there exist only very few investigations into pseudo communication situations or real noise (e.g. "multitalker babble" ([5]) and communication inside a jet cockpit ([8])). As already mentioned in an earlier study by Junqua et al., "studies of the Lombard reflex where data has been recorded while subjects are reading a list do not accurately represent the real conditions." ([9]).

One of the most frequently experienced noisy situations in the modern world is within the automobile. This presents a natural setting for verbal communication - be it with other passangers or by means of hands-free phone sets or even when talking to the car's dialog system. The goal of the present study is to test whether the Lombard reflex can be observed in an automotive environment and in dialog speech in the same way as in the laboratory with read speech. More specifically our hypotheses are:

1. The average and dynamic range of fundamental frequency (f0) increases with rising noise level.
2. The average and dynamic range of intensity increases with rising noise level.
3. Speaking rate decreases with rising noise level.

Additionally - and to our knowledge for the first time - we want to test for gender specific differences in the Lombard effect. Finally, all analyses should be carried out without manual annotation or segmentation to allow for automated Lombard detection based on our findings.

The remaining paper is structured as follows: The next section describes the recording of the speech material for the experiment. Section 3 describes the extracted features to test the above hypotheses as well as the statistical framework of our experiment. Section 4 presents the results while the last section contains discussion and future work.

2. Recorded Speech Data

We used parts of the German speech corpus HOESI which was recorded in 2008 by BAS Services, Munich, in cooperation with SIEMENS company, Munich. HOESI contains multi-channel dialog recordings of the same communication partners in sev-

[1] The rise in pitch is most likely a result of the rise in intensity.

eral automotive and lab conditions.[2]

The part relevant for this study consists of 24 speakers in total (12 women and 12 men), aged between 46 and 74 years. Speakers were recruited quite strictly: they had to be over 45 years old, were not dialect speakers and did not have any hearing or speaking impairments. The speakers were assigned into 12 same-gender pairs. Each pair produced four dialogs lasting about 7min under varying noise conditions (L0-L3), where L0 represented the quiet (neutral) condition and L1-L3 the noisy conditions with increasing noise-level. The noise was recorded from inside a standard car[3] moving at different velocities. Three constant speed levels were selected for each of the perturbing conditions: L1 = 80km/h, L2 = 120km/h and L3 = 160km/h. For each recording the speakers were placed in front of each other in a soundproof room and discussed a given topic in a free dialog; topics varied through Lombard conditions and ranged from discussions about politics to narrations about vacations or hobbies. During all Lombard conditions L1-L3 subjects wore prototypes of acustically closed behind-the-ear hearing aids (BTE) especially designed for this experiment. The BTEs allowed the impairing noise to be introduced directly into the ear, while at the same time the listener could hear the communication partner and his/her own voice without loss over the microphone of the BTE thus sustaining full auditory feedback. See Figure 1 for a detailed wiring diagram. The speech signal analysed for this study was not taken from the BTE input but rather recorded from two Beyerdynamic Opus 54 headset microphones carried by the speakers to minimize reverberation and cross-over to the noise channel; sampling rate was 48kHz and resolution 16bit. To ensure that both dialog partners uttered about the same amount of speech a test supervisor was present during the dialog and interfered by gesture if necessary.

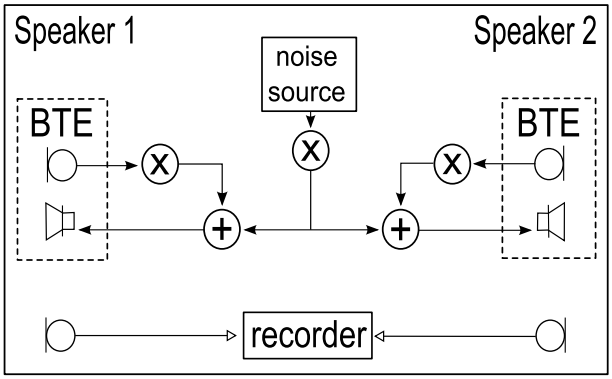

Figure 1: *Detailed wiring diagram for the recording. To achieve the correct noise levels gains (X) are adjusted by a psychoacoustic comparison inside the running car.*

3. Method

3.1. Analysed Features

The 96 recordings (24 speakers x 4 conditions L0-L3) were segmented into speech and non-speech using Praat[4]; only the parts tagged as speech were used for further analysis.

[2]Unfortunately the total HOESI corpus is not yet freely available due to copyright negotiations. If you are interested in working on these data, please contact the authors for possible access to parts of the corpus.

[3]Volkswagen Passat

[4]http://www.fon.hum.uva.nl/praat/, version 5.1.19 used.

F0 contours were calculated based on the headset microphone signal using the Schäfer-Vincent algorithm ([10]) and converted into semitones on a basis of 100Hz to make f0 dynamics comparable across genders. The Schäfer-Vincent algorithm marks non-voiced speech parts quite reliable; non-voiced parts were excluded from further f0 analysis.

We then calculated root-mean-square energy contours (RMS) using a Blackman window of 200ms size and 20ms shift. From these two contours a set of long-term features (LTF) as listed in Table 1 were calculated for each recording.

Table 1: *Long-term features derived from fundamental frequency and enegy contours.*

f0-m	median of f0
f0-r	range of f0 (quarter-quantile distance)
rms-m	median of RMS
rms-r	range of RMS (quarter-quantile distance)
rms-1	median of RMS differences between successive RMS minima and maxima
rms-2a	mean time distance of successive RMS maxima
rms-2b	mean time distance of successive RMS minima
rms-3a	median of RMS differences between mean and minimum
rms-3b	median of RMS differences between mean and maximum
rms-4a	median of slope of rising flanks
rms-4b	median of slope of falling flanks

Features rms-1 to rms-4b are called *rhythmicity features* based on a method developed by Chr. Heinrich (see [11] for details). In a nutshell, the RMS contour is first normalized to the average RMS of the total recording and then stylized into a succession of maxima and minima where maxima are defined as 'above average RMS' and minima as 'below average RMS'. Silence intervals greater than 1sec are filtered from this stylized energy contour and then the rhythmicity features as indicated in Figure 2 are extracted and averaged (or median) over the total recording. Aside from the means/medians as listed in Table 1 we also calculated the standard deviation/quarter-quantile distances, thus resulting in 18 measured LTF per recording and speaker.

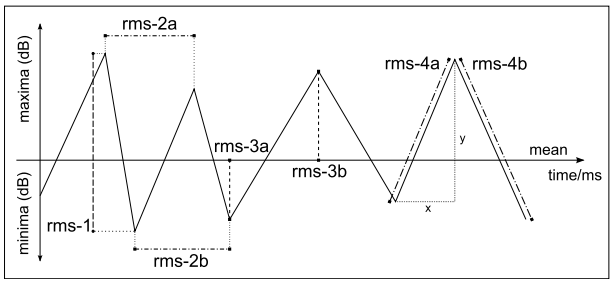

Figure 2: *Stylized RMS contour as basis for rhythmicity features.*

Note that the rhythmicity feature rms-1 is not the same as the quarter-quantile range of RMS rms-r: while rms-r is calculated from all absolute RMS contour values, the rms-1 feature is taken from the stylized (and normalized) contour max-

ima and minima values. It is therefore a more robust representation of the energy fluctuation typical for speech (CVCVC...) and more robust against varying recording conditions such as distance from mouth to microphone.

The rhythmicity features rms-2a/rms-2b reflect the regularity of minima/maxima succession and are therefore inversely correlated to speaking rate ([12]).

Features rms-3a/b can reveal a possible asymmetry of the stylized RMS contour, while rms-4a/b combine speaking rate and intrinsic dynamics.

3.2. Statistic

For all statistical analyses *mixed models* (MM) were applied ([13]). The four increasing Lombard conditions follow a trend which is testable for significance within a MM, which also allows the speaker to be treated as a random factor[5]. Speaker *gender* was treated as an additional between-speaker factor in the MM. F-values higher than the threshold of 8.49 are considered to be significant with a p-level of less than 0.01 ([14]). In post-hoc tests MMs were again applied, but only on split gender groups to test whether the factor *Lombard condition* has an influence on the features depending on the listener's gender.

4. Results and Discussion

Table 2 displays the results of the MM trend test for the individual features with regard to the four Lombard conditions. The medians of fundamental frequency *f0-m* and intensity *rms-m* increase significantly across the four Lombard conditions *L0-L3* (see Figure 3 and Figure 4); there is no significant gender difference *for the trend* on both of the main LTF.[6]

Table 2: *Mixed Models statistics for extracted features with regard to Lombard condition*

logogram	significance / interactions
f0-m	sig. (F=323.9, p<0.01)
f0-r	sig. (F=15.2, p<0.01) sig. interaction on gender: female (F=33.2, p<0.01) male n.s.
rms-m	sig. (F=316.5, p<0.01)
rms-r	n.s.
rms-1	sig. (F=20.4, p<0.01)
rms-2a	n.s.
rms-2b	n.s.
rms-3a	n.s.
rms-3b	sig. (F=27.9, p<0.01)
rms-4a	n.s.
rms-4b	sig. (F=26.2, p<0.01)

The range of fundamental frequency *f0-r* shows no significance on *gender* and *Lombard condition* but there is a significant interaction between these two factors. Therefore gender groups were tested separately which results in a significant trend

[5]To eliminate speaker-dependent idiosyncrasies, for instance f0 register.

[6]Of course there is a significant difference on absolute f0 between genders, but there is no significant difference between gender in the trend.

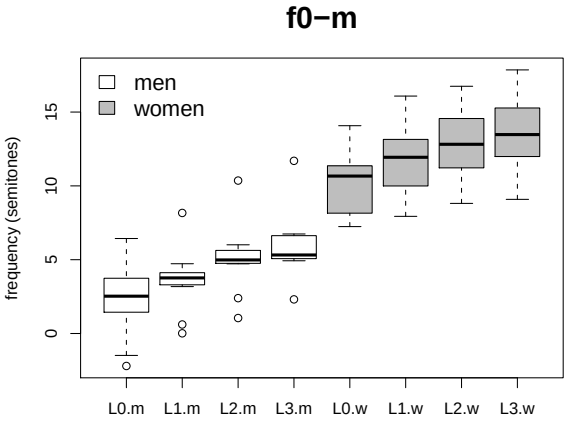

Figure 3: *Boxplots LTF median f0 across 4 Lombard conditions separated by gender.*

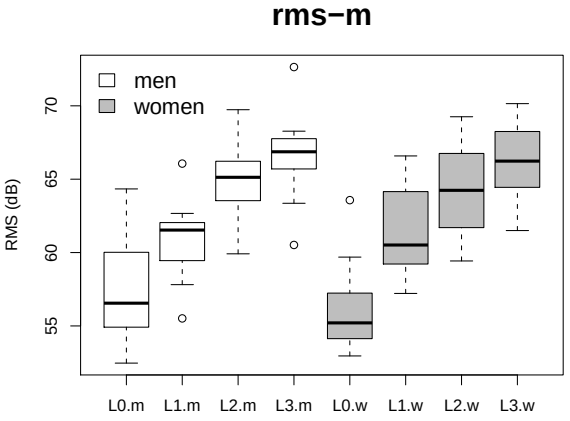

Figure 4: *Boxplots LTF median RMS across 4 Lombard conditions separated by gender.*

across *Lombard condition* for female speakers only (see Figure 5): while men keep their f0 range quite consistent, women extend it with every increased Lombard condition.

RMS range *rms-r* shows no significant difference across Lombard conditions.

So far, hypothesis 1 and partly hypothesis 2 can be confirmed: while the absolute LTF of fundamental frequency and intensity rises with increasing noise, the range basically remains constant except for the f0 range of female speakers.

On the other hand, the dynamic of the RMS increases with perturbing noise resulting in significant trends in some of the rhythmicity features *rms-1, rms-3b, rms-4b*. Looking more closely we find that this significance is again mainly due to the female speakers (*rms-1*: F=47.2, p<0.01, *rms-3b*: F=41.1, p<0.01). Men do not show an increase in their rhythmicity features, while female speakers do it to such an extent that the main factor for both genders is significant. The only exception is the feature *rms-4b* (slope of falling flanks) which is significant

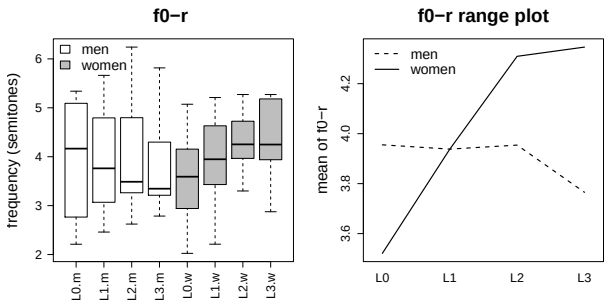

Figure 5: *Left: boxplots for f0 range f0-r throughout Lombard conditions. Right: range-plot for the mean of f0-r per Lombard condition separated into genders.*

across Lombard conditions for both genders.

Regarding the range of the rhythmicity features (not shown in Table 2) only *rms-1* was found to be significant across *Lombard conditions*, again only for female speakers.

Therefore, regarding the second part of hypothesis 2 – RMS range increases with Lombard noise – we have some mixed results. While the range of LTF RMS does not change, certain rhythmicity features representing intrinsic intensity dynamics differ across Lombard conditions, but only for the female speakers.

As for the timing features *rms-2a, rms-2b* there is no indication that subjects slow down their speaking rate with increasing Lombard noise.

Hence, the third hypothesis (and results of earlier studies regarding read lab speech) that speakers slow down when impaired by noise cannot be confirmed by our data.

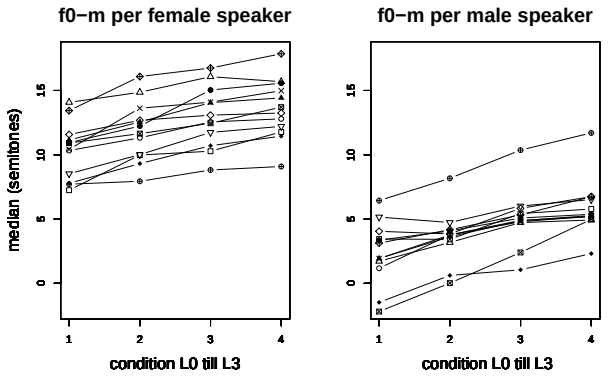

Figure 6: *F0 medians across Lombard conditions for each speaker; left: female speakers, right: male speakers.*

Another interesting observation regarding gender-specific behaviour was made on the LTF fundamental frequency. Figure 6 shows the trend of this feature for each speaker, separated for female (left) and male (right) speakers. With increasing Lombard noise the fundamental frequency of the majority of the male speakers is concentrated in a small range, whereas the female speakers maintain their individual register throughout Lombard conditions. Whether this covergence is typical for male social behaviour or just a random effect remains to be clarified.

5. Conclusion

We studied spontaneous dialog speech of 24 native German speakers under real noise Lombard conditions to verify earlier findings about the Lombard reflex. The recordings involved a sophisticated experimental setting where natural hearing was replaced with modern hearing aids. This allowed natural noise recorded from an automotive environment to be inserted into the auditory feedback loop.

Our results show that the main Lombard effects, rising of intensity and fundamental frequency, can be confirmed for both genders. The range of fundamental frequency was increased as well, but only for the female speakers. Long term range of intensity did not change across Lombard conditions, but intrinsic rhythmicity features representing the dynamics of root mean square energy did - again only for female speakers. Contrary to our expectation speaking rate did not slow down with Lombard noise.

In conclusion we can say that dialog processing in noisy environments will have to deal with the same Lombard effects known for read speech, except for speaking rate. In addition, Lombard effects cannot be expected to be constant across genders.

6. References

[1] Lombard É (1911) Le signe de l'élévation de la voix. In: Ann. Mal. Oreil. Larynx, Vol 37, pp. 101-119.

[2] Lane H, Tranel B (1971): The Lombard sign and the role of hearing in speech. In: Journal of Speech and Hearing Research, Vol 14, pp. 677-709.

[3] Charlip W, Burk K (1969): Effects of noise on selected speech parameters. In: Journal of Communication Disorders, Vol 2, pp. 212-219.

[4] Junqua J-C, Fincke S, Field K (1999): The Lombard Effect: a reflex to better communicate with others in noise. In: Proc. of the ICSLP, pp. 2083-2086.

[5] Davis C, Jeesun K, Grauwinkel K, Mixdorff H (2006): Lombard Speech: Auditory (A), visual (V) and AV effects. In: Proc. of Speech Prosody, pp. 361-365.

[6] Lindblom B (1990): Explaining phonetic variation: A sketch of the H and H theory. In: Hardcastle W & Marchal A (eds.): Speech Production and Speech Modeling. Dordrecht a.o.: Kluwer Academic Publishers, pp. 403-439.

[7] Uchanski R, Choi S, Braida L, Reed C, Durlach N (1996): Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. In: Journal of Speech and Hearing Research, Vol 39, pp. 494-509.

[8] Stanton B, Jamieson L, Allen G (1988): Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions. In: Proc. of the ICASSP, pp. 331-334.

[9] Junqua J-C, Fincke S, Field K (1998): Influence of the speaking style and the noise spectral tilt on the Lombard reflex and automatic speech recognition. In: Proc. of the ICSLP, pp. 467-470.

[10] Schäfer-Vincent K (1983): Pitch period detection and chaining: method and evaluation. In: Phonetica, Vol 40, pp. 177-202.

[11] Schiel F, Heinrich Chr, Neumeyer V (2010): Rhythm and Formant Features for Automatic Alcohol Detection. In: Proc. of the INTERSPEECH, Chiba, Japan, pp. 458-461.

[12] Heinrich Chr, Schiel F (2011): Estimating Speaking Rate by Means of Rhythmicity Parameters. In: Proc. of the INTERSPEECH, Florence, Italy, submitted.

[13] Baayen R H (2008): Analyzing Linguistics Data: a practical introduction to statistics. Cambridge University Press.

[14] Reubold U, Harrington J, Kleber F (2010): Vocal aging effects on F0 and the first formant: a longitudinal analysis in adult speakers. In: Speech Communication, Vol 52, pp. 638-651.