How Does Lombard Speech Behave When The Background Noise Level Decreases?

Paul Dräger[1], Jochen Steffens[2], Markus von Berg[3]

[1] *Institute of Sound and Vibration Engineering, 40476 Düsseldorf, paul.draeger@web.de*
[2] *Institute of Sound and Vibration Engineering, 40476 Düsseldorf, jochen.steffens@hs-duesseldorf.de*
[3] *Institute of Sound and Vibration Engineering, 40476 Düsseldorf, markusmartin.vonberg@hs-duesseldorf.de*

Introduction

The Lombard effect is a well-researched phenomenon that has been studied by different scientific disciplines. The effect is named after the surgeon and otorhinolaryngologist Étienne Lombard. In his work published in 1911, he describes that his patients, apparently unconsciously, raise their speaking level when exposed to background noise [1].

The changes in human speech production due to the Lombard effect are commonly referred to as 'Lombard speech'. Previous research has found ratios between 0.1 dB and 0.38 dB increase in speaking level per 1 dB additional level of ambient noise [2-4]. However, these 'Lombard slopes' depend on the type and composition of the ambient noise [4;5]. Lombard speech is also characterised by changes in speech production, such as an increase in the fundamental frequency (F0) and the first formant (F1) [6-8]. Regarding the shift of the second formant (F2), previous results are not consistent as for F0 and F1 [9]. However, these spectral changes are accompanied by a general 'smoothing of the spectral slope' [7].

Furthermore, Shen, Cooke and Janse [10] observed that the impact of the Lombard effect on speaking behaviour also varies over time, and that the properties of Lombard speech are not static, but dynamic. This observation raises further questions about the temporal changes in both the Lombard effect and the production of Lombard speech.

This study investigates how typical features of Lombard speech behave when the amplitude of the background noise is reduced. To that end, Lombard speech was triggered by a background noise that was presented at 70 dB(A) and then continuously reduced to 50 dB(A) over a period of 100 seconds. The speech characteristics during this 'post-fade' phase were compared to speech during a 'baseline'-measurement with a steady background noise at 50 dB(A).

It was expected that, due to the previous exposure with louder background noise, participants would exhibit stronger characteristics of Lombard speech during the 'post-fade' phase than the characteristics during the 'baseline' condition. 50 dB(A) was chosen as the baseline level to enable an unnoticed reduction in the level, whereas a reduction to 0 dB(A) would have drawn to much attention to the noise level completely vanishing. And since previous work found sound levels of 40-50 dB to be the approximate 'starting threshold' for the Lombard effect [11;12], a level of 50 dB(A) seemed to be a suitable baseline level.

Methods

Sample

The experiment was conducted with 34 participants (f: 16; m: 18; d: 0) between 20 and 67 years (median 23 years, SD: 10,6), who held a conversation in teams of two. In order to create a comfortable situation for the conversation, participants were recruited in pairs, and most teams already knew each other before the experiment. Participants were not informed on the subject and the research question before taking part in the experiment. The only prerequisite for participation was the self-report of no known hearing impairments.

Material

Since the Lombard effect is more pronounced in communication situations [13], the experiment aimed at evoking a natural conversation between the two members of a team, to assure natural language production by the participants. On the other hand, the individual conversations should be similar as possible to allow valid comparisons between teams. The 'DiapixUK task' [14] fulfils both requirements for this experiment. In this task, both dialogue partners receive two versions of a picture differing in 12 'errors'. The task of the teams is to find these errors without being able to see the other person's version of the picture. The corpus of the 'DiapixUK task' consists of 12 pictures with two versions each. In this experiment, four pictures per team were required for the given lengths of background noise presentation (see 'Procedure' section).

Regarding acoustical stimuli, a 'babble noise' was created by layering duplicates of 12 recordings of different speakers. These duplicates were cut and rearranged in time to create a sound that was as homogeneous as possible. However, as the result contained a unpleasant amount of high-frequency energy, which lead to unpleasantly harsh sibilants, a more natural sound was created using a second-order low-pass filter (Adobe Audition parametric equaliser, cut-off frequency: 5670 Hertz) and a de-esser (Izotope RX9 De-Ess, cut-off frequency: 5860, threshold: -40 dB).

Design and procedure

The test subjects were seated in two different rooms, but could see each other through a window. They spoke and listened to the other team member using Beyerdynamic DT-297s headsets. To compensate the headphones' attenuation of the airborne sound [15], participants also heard their own voice through the headsets. The audio connection was realised via a Dante network and audio interfaces (ESI Planet 22x), and the background noise level was calibrated using a dummy head (Head Acoustics HMS IV.0 with BEQ II preamplifier).

For better readability, the two members of a team will be referred to as 'person A' and 'person B' in the remainder of this paper. In the first round of a condition, person A was assigned the 'active partner' (i.e., the person should describe their image). As the 'passive partner', person B was asked to compare this description to their own picture and respond accordingly. This procedure was chosen to ensure that one participant had a high proportion of the conversation per round. In the next iteration, the pictures were changed and the roles swapped, so that person B was now the active partner and person A the

passive partner. The experiment was divided into two conditions with two such iterations each. In the first 'baseline' condition, the babble noise was presented at a sound pressure level of 50 dB(A). The two iterations of this condition lasted 3 minutes each. In the 'fade' condition, the noise level was at 70 dB(A) for the first 100 seconds, followed by a 100-second-fade to 50 dB(A), and a further 100 seconds with a level of 50 dB(A). Participants were given a sample picture prior to the experiment to introduce them to the task's paradigm. The selection of the pictures was randomised across all conditions and iterations to eliminate picture-dependent influences. The order of the conditions was determined in such a way that the 'baseline' condition was always presented before the 'fade' condition to prevent possible impacts of previously presented louder background noise on the speech production during the 'baseline' condition.

Analysis

In the analysis, the speech level, the fundamental frequency, the first two formants and the spectral rolloff of the audio voice recordings of person A were compared across conditions. The majority of the pre-processing was carried out using Python scripts. Among others, the Python interface 'Parselmouth' of the phonetic analysis programme 'Praat' was used for this purpose. The data obtained was then transferred to the programme 'Jamovi' (v2.4.1) for statistical analysis. The speech recordings during the 'fade' condition was subdivided into three phases according to the changing noise level:

- 'pre-fade': [0:00 – 1:20 min]
- 'during-fade': [1:20 – 3:00 min]
- 'post-fade': [3:00 – 4:40 min]

Linear mixed-effects (LME) models were then calculated to predict each feature by the condition (or phase) as a fixed effect and the participant as a random intercept. Subsequent post-hoc tests were used to determine the differences magnitude and statistical significance between single conditions.

Results

Changes in level and F1 during the 'fade' condition

Figure 1 displays the changes in speech sound level in the 'fade' condition. To calculate the intensity values, the pauses in speech were interpolated and then average values were calculated from the individual frames of all speech recordings.
As depicted by Figure 1, the highest sound level appears in the 'pre-fade' phase, when the babble noise level is at 70 dB(A). This phase is followed by a reduction in the level as the babble noise level decreases until the lowest speech levels are reached in the 'post-fade' segment. Figure 2 illustrates the respective changes in the first formant F1, which, on the other hand, shows no reduction in the values in the 'post-fade' phase.

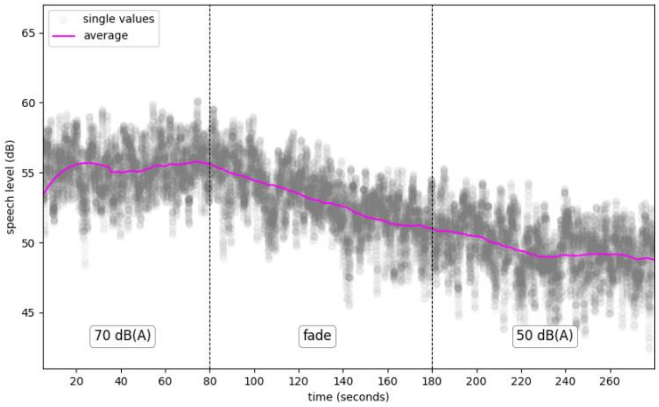

Figure 1: speech level in the 'fade' condition. Pauses were replaced by interpolated values.

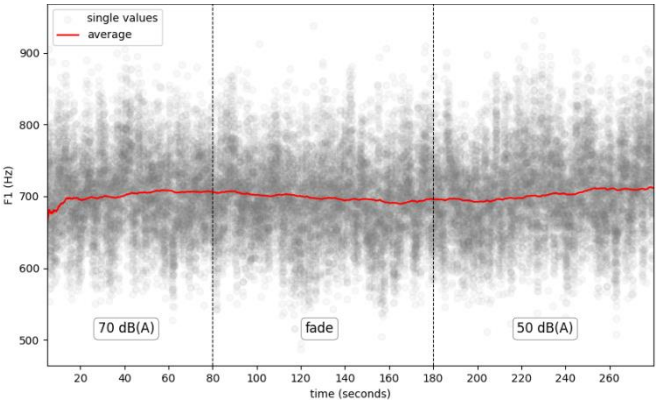

Figure 2: F1 values in the 'fade' condition.

Comparison of 'baseline' and 'pre-fade' condition

First, the increase in Lombard speech during the 'pre-fade' phase of the 'fade' condition was investigated by comparing the average values of the extracted audio features to those in the 'baseline' condition. The post-hoc tests of the LME model showed that the average speech level recorded during the 'baseline' condition was 5 dB lower than in the 'pre-fade' phase ($p < 0.01$; Cohen's $d = -1.603$). The marginal R^2 (R^2m) equalled 0.393, meaning that almost 40 % in speech level variance could be explained by the different conditions (or phases). The conditional R^2 (R^2c) constituting the variance explained by both the fixed and random effect was at 0.798.
The fundamental frequency F0 was on average 14 Hz lower during the 'baseline' condition ($p < 0.01$; $d = -0.379$), and the model revealed a small variance explanation of $R^2m = 0.035$ ($R^2c = 0.940$).
The frequencies of F1 in the 'baseline' condition were on average 23 Hz below the F1 values in the 'pre-fade' segment ($p < 0.01$; $d = -0.227$), with $R^2m = 0.013$ ($R^2c = 0.972$).
The second formant F2 is on average 19 Hertz lower in the 'baseline' condition than in the 'pre-fade' segment ($p = 0.001$; $d = -0.116$), and there was a statistically significant effect, however, again at a small R^2m of 0.003 ($R^2c = 0.982$). The limit for the spectral rolloff was set to 90% of the cumulative energy. The average difference in spectral rolloff between the 'pre-fade' and the 'baseline' condition was 305 Hz ($p < 0.01$; $d = -0.819$) and R^2m was at 0.146 ($R^2c = 0.854$).

Comparison of 'baseline' and 'post-fade' condition

In the next step, the mean values of the 'baseline' condition were compared to those of the 'post-fade' phase in the 'fade'

condition in the time interval [3:00 - 4:40 min]. Here, the babble noise was at 50 dB(A) in both conditions. The speech level was 0.33 dB lower in the 'baseline' condition, but the difference was not statistically significant ($p = 0.458$; $d = -0.1$, $R^2m = 0.002$, $R^2c = 0.705$).

Furthermore, F0 was on average 0.22 Hz higher during the 'baseline' condition, which was again not statistically significant ($p = 0.857$, $d = 0.004$) and there virtually no variance explained by the condition ($R^2m < 0.001$, $R^2c = 0.985$). By contrast, the values of F1 were significantly lower in the 'baseline' condition ($p < 0.01$; $d = -0.22$), and 1,2% of the variance was explained by the condition ($R^2m = 0.012$, $R^2c = 0.951$).

No significant difference was observed for F2, which was on average 3 Hz lower in the 'baseline' condition ($p = 0.678$; $d = -0.016$, $R^2m < 0.001$, $R^2c = 0.974$).

Finally, the average difference in the spectral rolloff between the 'baseline' condition and the 'post-fade' phase was at −42 Hz and the p-value slightly missed statistical significance ($p < 0.075$; $d = -0.133$, $R^2m = 0.004$, $R^2c = 0.911$).

Tab. 1: Descriptive statistics across conditions

	Condition	Int.	F0	F1	F2	SRO_90
Mean	Baseline	48,9	171,1	682,8	1534,6	896,2
	Pre-fade	54,1	185,4	706,2	1554,1	1200,9
	Post-fade	49,3	170,9	705,2	1537,3	937,7
SD	Baseline	3,6	40,4	98,9	162,8	316,6
	Pre-fade	2,7	34,8	106,3	171,2	419,9
	Post-fade	3,1	40,3	104,6	170,4	305,8

Tab. 2: Post-hoc tests for significant differences between conditions

	Baseline-Prefade	p	Baseline-Post-fade	p
Level	-5,129	< 0,01	-0,33	0,458
F0	-14,34	< 0,01	0,22	0,857
F1	-23,41	< 0,01	-22,4	< 0,01
F2	-19,48	0,001	-2,73	0,678
Spectral rolloff	-304,7	<0,01	-41,5	0,075

Discussion

This study analysed the characteristics of Lombard speech the event of a slowly decreasing background noise level. The levels observed in the 'baseline' condition with a background noise level of 50 dB(A) were compared to those of the 'pre-fade' phase and the subsequent 'post-fade' phase of the second 'fade' condition, where the background noise level was at 70 dB(A) and 50 dB(A), respectively. The hypothesis stated that due to the previous exposure to a louder masker in the 'pre-fade' phase, the exhibition of Lombard speech during the 'post-fade' segment would be stronger than in the 'baseline' condition.

Considering the differences between the 'pre-fade' phase (noise level: 70 dB(A)) and the 'baseline' condition (noise level: 50 dB(A)), a more pronounced Lombard effect could be observed for all extracted audio features. The intensity of the speech recordings was on average approximately 5 dB higher than with the reference condition. With a noise level difference of 20 dB, the resulting Lombard slope of 0.25 dB level increase per 1 dB additional background noise level was within the range of other studies [2-4]. The other audio values also showed the expectable differences for the given increase in noise level.

Contrary to the high values of the conditional R^2c, linear mixed-effects models showed small values for the marginal R^2. This suggests that the variance is well explained by the models overall, predominantly by differences across participants, but the influence of the condition was rather small.

The comparison of the 'post-fade' phase with the reference condition 'baseline' (noise level 50 dB(A) in both cases) only showed a small insignificant difference. Similarly, non-significant results were obtained when analysing the fundamental frequency (F0) and the second formant F2. The spectral rolloff provides information about the spectral composition and the shift of energy into higher frequency ranges. Although the frequency for the 'post-fade' segment is on average 41 Hertz higher than in the 'baseline' condition indicating a more pronounced Lombard speech, this finding was also not significant. However, the influence of the comparatively small sample size on the p-values has to be taken into account for all analyses, and despite the predominantly statistically non-significant results, tendencies can be recognised. On the contrary, the difference in F1 is statistically significant, and it is on average 22 Hertz higher in the 'post-fade'. While the temporal changes of the other audio features decrease, i.e. an approach to the values of the reference condition, the course of the first formant F1 does not. Future research might investigate this phenomenon and its origin using different experimental conditions.

Considering the later part of an ongoing conversation, the test paradigm applied here has a potential shortcoming: Although no team actually found all 12 errors in the pictures of the 'DiaPixUK task, most of the conversations tended to 'stagnate' at the end. This mainly affected the closing 'post-fade' phase in the 'fade' condition. The 'baseline' condition was considerably shorter (three minutes compared to 4:40 minutes 'fade' condition) and therefore, less affected by this phenomenon. Removing the last seconds of the 'post-fade' phase from the analysis did not show any considerable influence on the results of the statistical analysis. However, the interruption affects the representation of the time sequences, as all recordings had to have the same length for the analysis, and pauses in speech were not removed, but interpolated. An influence of this interpolation on the temporal changes on the audio features cannot be excluded. For this reason, this temporal envelope of each feature was only used for illustrating purposes (Figure 1 and 2), while the statistical analysis were based on the temporal mean values of each phase. As mentioned above, the number of errors in the pictures was not the issue for the fading conversation, but the limitation of areas that seem likely to contain errors. Giving more than one pair of pictures to participants could contribute to a more homogeneous 'conversation intensity' across all conditions and phase. Also, different methods to enforce speech production could be applied. For example, having texts read aloud could be more effective means to ensure continuous recordings of speech. Such a task eliminates the communicative intention which also affects the intensity of Lombard speech, but it does not require the presence of a second participants also removes potential effects of co-dependency between the two participants. Furthermore, both the duration of background noise exposure and the noise level could be extended in subsequent experiments, to create a more realistic situation for the use of Lombard speech. In this experiment, the higher noise level

was presented for duration of 80 seconds, to design an experiment of acceptable test length. However, in situations where the Lombard effect typically occurs, such as in restaurants or vivid streets, the time of exposure can be several minutes or even hours. A field study might be a suitable methodological approach to investigate the decrease in Lombard speech after longer periods of noise exposure.

Conclusion

It was investigated whether exposure to a loud background noise and subsequent lowering of the sound pressure level would result in more pronounced Lombard speech than in the case of no previous exposure to louder noise. While most characteristics of Lombard speech did not support the concept of a sustaining 'hysteresis', F1 hardly changed after decreasing the noise level. To what extend these findings apply to different experimental conditions with longer exposure times or different noise levels, has to be investigated in future research.

References

[1] Lombard, É.: Le signe de l'élévation de la voix. Annals Maladiers Oreille Larynx, Nez Pharynx (37), S. 101–119, 1911.

[2] Dreher, J. J.; O'Neill, J.: Effects of ambient noise on speaker intelligibility for words and phrases. Journal of the Acoustical Society of America 29 (12), S. 1320–1323, 1957.

[3] Korn, T. S.: Effect of psychological feedback on conversational noise reduction in rooms. Journal of the Acoustical Society of America 26 (5), S. 793–794, 1954.

[4] Stowe, L. M.; Golob, E. J.: Evidence that the Lombard effect is frequency-specific in humans. Journal of the Acoustical Society of America 134 (1), S. 640–647, 2013.

[5] Garnier, M.; Bailly, L.; Dohen, M.; Welby, P.; Lœvenbruck, H.: An Acoustic and Articulatory Study of Lombard Speech: Global Effects on the Utterance. In: International Speech Communication Association (Hg.). INTERSPEECH 2006, 9th International Conference on Spoken Language Processing, Pittsburgh, S. 2246–2249, 2006.

[6] Godoy, E.; Koutsigiannaki, M.; Stylianou, Y.: Approaching speech intelligibility enhancement with inspiration from Lombard and Clear speaking styles. Computer Speech and Language 28 (2), S. 629–647, 2014.

[7] Lu, Y.; Cooke, M.: The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. Speech Communication 51 (12), S. 1253–1662, 2009.

[8] Van Summers, W.; Pisoni, D. B.; Bernacki, R. H.; Pedlow, R. I.; Stokes, M. A: Effects of noise on speech production: acoustic and perceptual analyses. Journal of the Acoustical Society of America 84 (3), S. 917–928, 1988.

[9] Maheswari, U.; Shahina, A.; Nayeemulla Khan, A.: Understanding Lombard speech: a review of compensation techniques towards improving speech based recognition systems. Artificial Intelligence Review 54, S. 2495–2523, 2021.

[10] Shen, C.; Cooke, M.; Janse, E.: Speaking in the presence of noise: Consistency of acoustic properties in clear-Lombard speech over time. The Journal of the Acoustical Society of America 153 (4), S, 2165, 2023.

[11] Bottalico, P.; Passione, I. I.; Graetzer, S.; Hunter, E. J.: Evaluation of the starting point of the Lombard effect. Acta Acustica United With Acustica 103 (1), S. 169–172, 2017.

[12] Lazarus, H.: Prediction of Verbal Communication in Noise – A review: Part 1. Applied Acoustics 19 (6), S. 439–464, 1986.

[13] Amazi, D. K.; Garber, S.: The Lombard sign as a function of age and task. Journal of Speech, Language and Hearing Research 25 (4), S. 581–585, 1982.

[14] Baker, R.; Hazan, V.: DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. Behavior research methods 43, S. 761–770, 2011.

[15] Garnier, M.; Henrich, N.; Dubois, D.: Influence of Sound Immersion and Communicative Interaction on the Lombard Effect. Journal of Speech, Language and Hearing Research 53, S. 588–608, 2010.