THE EFFECT OF LOMBARD SPEECH MODIFICATIONS IN DIFFERENT INFORMATION DENSITY CONTEXTS

Omnia Ibrahim, Ivan Yuen, Marjolein van Os, Bistra Andreeva, Bernd Möbius

Language Science and Technology, Saarland University
omnia@lst.uni-saarland.de

Abstract: Speakers adapt their speech to increase clarity in the presence of background noise (Lombard speech) [1, 2]. However, they also modify their speech to be efficient by shortening word duration in more predictable contexts [3]. To meet these two communicative functions, speakers will attempt to resolve any conflicting communicative demands. The present study focuses on how this can be resolved in the acoustic domain. A total of 1520 target CV syllables were annotated and analysed from 38 German speakers in 2 white-noise (no noise vs. -10 dB SNR) and 2 surprisal (H vs. L) contexts. Median fundamental frequency (F0), intensity range, and syllable duration were extracted. Our results revealed effects of both noise and surprisal on syllable duration and intensity range, but only an effect of noise on F0. This might suggest redundant (multi-dimensional) acoustic coding in Lombard speech modification, but not so in surprisal modification.

1 Introduction

It is known that exposure to background noise will trigger speech adaptation, also known as 'Lombard speech' [2]. Its goal is to enhance communication in challenging and degraded interactive situations. Such speech adaptation often takes the form of increased loudness, higher pitch, expanded vowel space, hyper-articulation, and lengthening [4, 5, 6], with perceptual consequences of improved intelligibility [7, 8].

At the same time, speakers are reported to reserve effort and adopt an efficient communication strategy. Therefore, they tend to produce more reduced forms or shorter durations for predictable/probable messages. This predictability effect reflects a tendency towards communicative efficiency, when the sender and receiver can achieve successful communication with minimal effort on average in a reliable channel. There is a growing body of research suggesting the pervasiveness of predictability in language, as manifested in phonetic reduction of words that have higher n-gram probabilities [3, 9, 10], or choices of linguistic/syntactic units [11]. This entails that speaker's choices and listener's preferences are affected by the probability and frequency of occurrence of such units. There are several measures to quantify the amount of information conveyed in a message [12]. One of them is surprisal. Surprisal captures the intuition that linguistic expressions that are highly predictable in a given context convey less information than those that are unexpected. surprisal is defined as the contextual predictability of a unit and can be used as a measure of the amount of information that is conveyed by that unit in terms of bits, using Equation (1) where S stands for surprisal and P for probability:

$$S(unit_i) = -log_2 P(unit_i|Context) \tag{1}$$

However, if the channel is unreliable, for instance in a noisy environment, the speaker will need to reconsider the advantages and disadvantages of efficiency in the context of predictability

[13]. In an unreliable channel, an efficient coding of a message might potentially increase the likelihood of being mis-heard and mis-understood. The primary goal of this study is to investigate the effects of background noise and contextual predictability (defined as surprisal) on three features extracted on the syllable level: duration, intensity, and fundamental frequency (F0). We tested whether an unreliable noisy channel will trigger additive or interactive effects on speech enhancement arising from surprisal. We expected that Lombard speech and surprisal would interactively affect our acoustic variables, because speakers will enhance high surprisal units more than low surprisal units in noise condition. This strategy will allow speakers to be both informative and efficient.

2 Methods

2.1 Participants

Thirty-eight native German speakers were recruited (12 M, 26 F in the range of 19-60 years), with no known hearing and speaking impairments. We adopted the relatively wide age range to include more participants, on the basis of the observation that young adults and older adults with age-typical mild high-frequency hearing loss do not differ in how they adjust their speech when speaking in the Lombard condition [14].

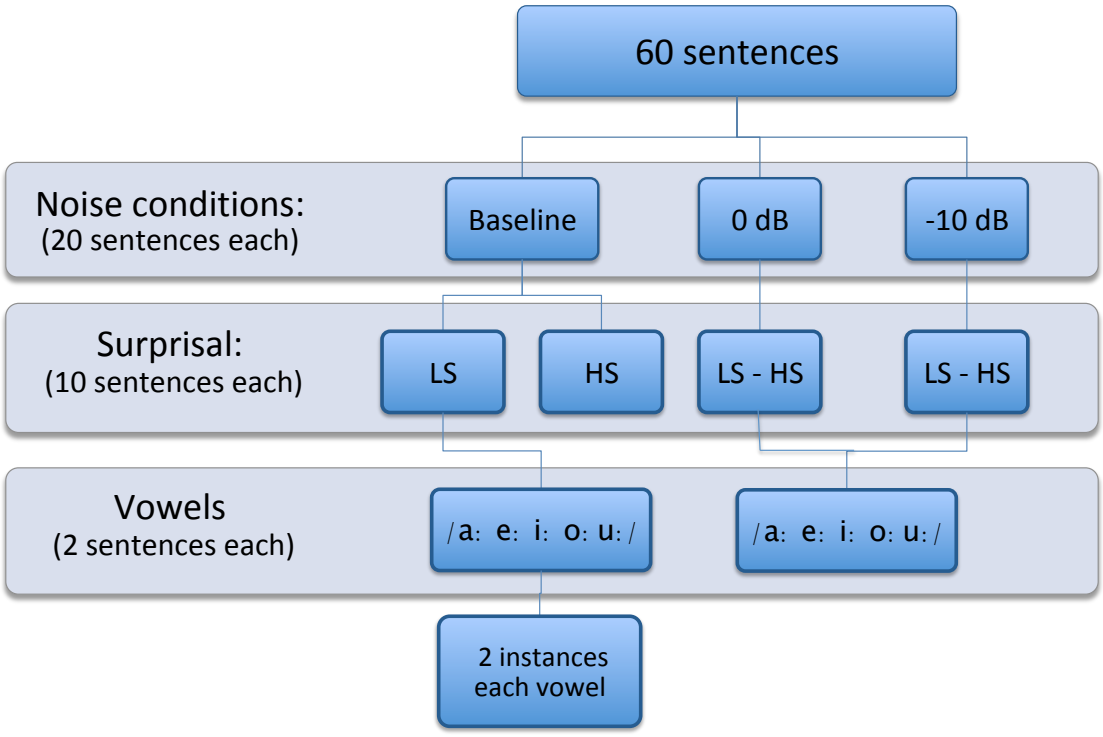

Figure 1 – Conditions and structure of experimental stimuli, constraining the selection of 60 sentences from DeWaC.

2.2 Material / stimuli

A set of 60 sentences was selected from a version of the DeWaC corpus [15, 16] based on high vs. low surprisal bins (HS vs. LS). The information-theoretic factor surprisal was estimated by means of a trigram syllable-level language model trained on DeWaC. The selected sentences were shortened so that they could be displayed on two lines on the screen while preserving the target syllable's surprisal value. Each target CV syllable was part of a polysyllabic word in a sentence, where C beginning with a /p, b, d, k/ plosive was combined with 5 vowels (/a:, e:, i:, o:, u:/). The target syllables occurred in three white-noise conditions (baseline= no noise, 0 dB and -10 dB SNR). The combination of 5 vowels in 2 surprisal contexts in 3 noise conditions with 2 repetitions resulted in a total of 60 stimuli, randomized for presentation (Figure 1).

2.3 Experimental procedures

The data were recorded in a soundproof booth. Speakers wore a DPA 4067-F Omni headset microphone to record the speech signal and AKG K271 MKII over-ear headphones to hear the white noise signal during the recording session. The stimuli were visually presented on screen as a slide (black font, white background) one at a time. A research assistant remotely controlled the advancement of the stimuli outside the recording booth. A practice session was provided for speakers before testing. During practice, speakers were instructed to read a different set of German sentences and the research assistant calibrated the equipment. The test phase consisted of 3 blocks. Speakers were informed about the presence of background noise in the first and last blocks. They stood upright and read the stimuli using their habitual reading pace. The order of the noise conditions (0 dB vs. 10 dB) was counterbalanced, retaining the middle block for the no-noise condition (baseline). This design was chosen to reduce the tendency of speakers to linearly adjust to noise. Productions were recorded and stored as a mono .wav file with a sampling rate of 48 kHz and 24 bits per sample.

2.4 Data annotation

The target words/syllables/segments were manually annotated by two student assistants (a native speaker of Vietnamese and a native speaker of German) using Praat [17], while the non-target words/syllables/segments in each sentence were automatically annotated using WebMAUS [18] (Figure 2).

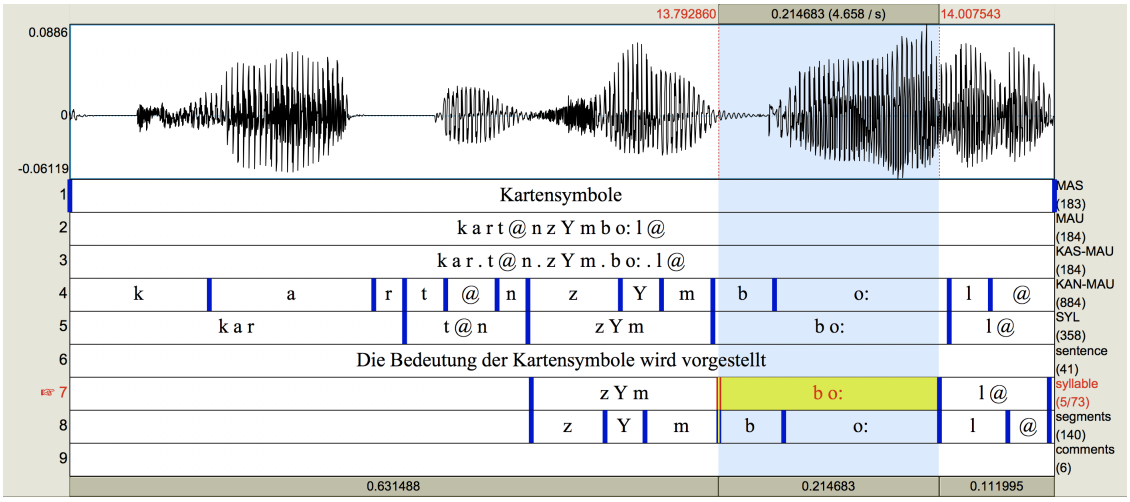

Figure 2 – Annotation example: automatic (tiers 1–5) and manual (tiers 6–9) annotation.

2.5 Data analysis

Three acoustic features were extracted from the target syllable (the focus of this paper) using in-house Python and Praat scripts: fundamental frequency (median), intensity, and total syllable duration. These features were chosen for the following reasons: Duration and F0 have previously been shown to be sensitive to predictability, and intensity to the presence of background noise (Lombard speech). The current analysis focused only on two conditions (baseline and -10 dB SNR), because these two conditions lie at the two ends of the noise continuum and should provide the clearest effect, if there is any. Linear mixed-effects modelling was used to evaluate the hypothesized effect(s) and interaction of noise and surprisal using R lmer package [19]. Backwards model selection procedure was applied to arrive at a final model as reported below. According to such procedure, a maximal random structure was first formulated to identify the model that best fit our data [20]. We included random intercepts and random slopes for all fixed effects. Our fixed effects are noise condition, surprisal group and vowel, while the random effects are speakers, gender, syllable, and stimulus. Our three dependent variables were z-scored by speaker, which minimized any pitch, intensity and duration differences between speakers. In case of convergence errors we reduced the maximal random structure step-wise. First, we removed random slopes, and then, if necessary, random intercepts. Significance of fixed effects was evaluated by performing maximum likelihood t-tests using Satterthwaite approximations to degrees of freedom.

3 Results

Syllable duration: The final model structure was as follows: *lmer(Dur-Syl-z ~ NoiseLevel + Surprisal-Group + Vowel + (1 |Stimulus) + (1 |Syllable))*.

As expected, there were main effects of noise (Estimate=.10491, t=2.994, p=.00281**) and surprisal (Estimate=.33927,t=2.138, p=.03821 *) on syllable duration. No other effects or interactions reached statistical significance.

Intensity range: The final fixed and random effect estimates for intensity range were as follows: *lmer(RangeInt-Syl-z ~ NoiseLevel + Surprisal-Group + Vowel + (1 | Stimulus) + (1 | Syllable))*. Similar to the results on syllable duration, there were main effects of noise (Estimate=1.712e-01, t=4.272, p=2.09e-05 ***) and surprisal (Estimate=3.717e-01, t=2.290, p=.0267 *) on intensity range. No other effects or interactions reached statistical significance (Figure 3).

Fundamental frequency: The final F0 model was as follows: *lmer(F0MedianSyl-z ~ NoiseLevel + Surprisal-Group + Vowel + (1 | Stimulus) + (1 | Syllable))*. Like syllable duration and intensity range, there was a main effect of noise (Estimate=.27661, t=6.962, p=5.5e-12 ***) on F0. However, no effect of surprisal was found (Estimate= 0.25139, t= 1.540, p= 0.12959), even though there was a trend for F0 to be higher in the -10 dB noise condition than in the baseline condition. No other effects or interactions reached statistical significance.

4 Discussion and conclusion

In the information-theoretic framework, predictability effects can manifest by reducing signal redundancy (i.e., source coding) to communicate efficiently: predictable words tend to be reduced and short in duration. On the other hand, noise effects can manifest by preserving redundancy (i.e., channel coding) to minimize errors: in noise conditions, loudness and fundamental frequency increase. Our question was how channel-coding modification, arising from Lombard speech, affects source-coding changes, arising from contextual predictability.

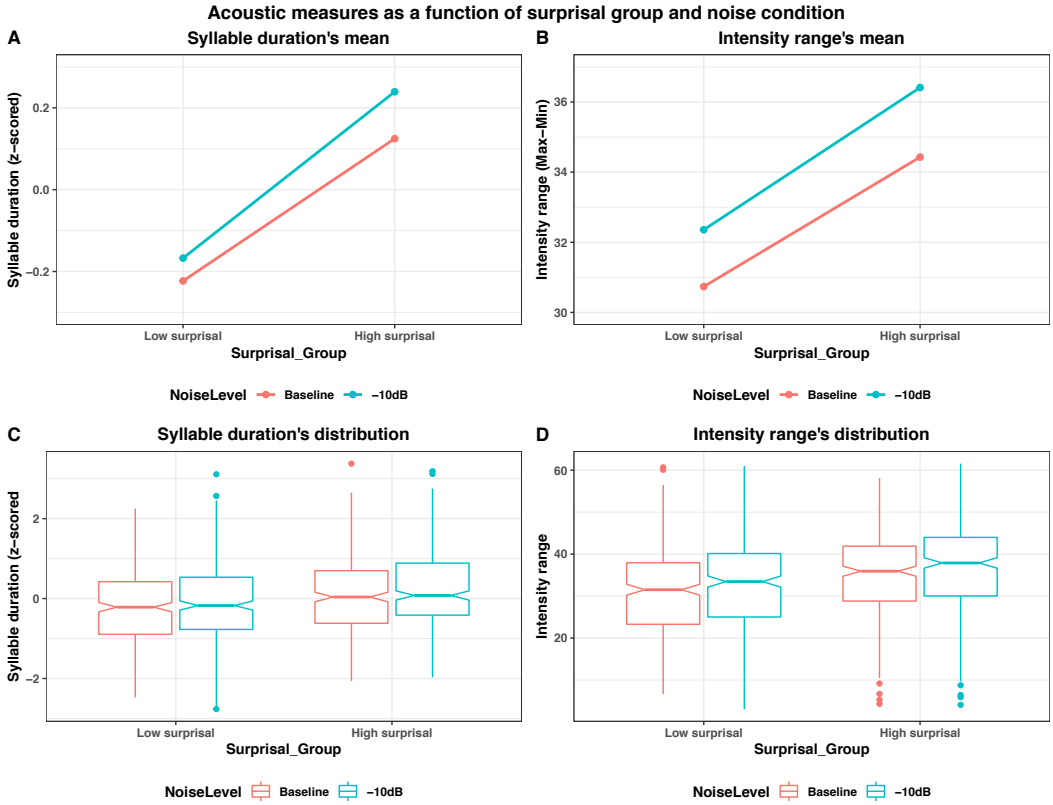

Figure 3 – Acoustic measures (A/C: syllable duration, B/D: intensity range) as a function of surprisal (low vs. high) and noise (red: baseline, blue: -10 dB)

Our results indicated no interaction between noise and surprisal conditions for any of the three acoustic features under study, counter to our prediction. This suggests that speech is enhanced additively as a function of noise and surprisal in intensity range and duration, but not in F0. Since we did not code the accent type on the target word, it remains unclear whether surprisal affects F0 at all. Interestingly, noise condition affected all three acoustic variables measured in this study, but surprisal condition didn't affect F0. From an information-theoretic perspective, the strategy of enhancing more than one acoustic variable in noise condition suggests that speakers opt for coding redundant acoustic signals to minimize errors in an unreliable channel.

Our findings are broadly consistent with many previous studies on the effects of background noise on the acoustic variables under study here. The results showed that syllable duration and intensity increased in noisy conditions for both high-surprisal and low-surprisal syllables, but there was no overall increase in fundamental frequency in noisy conditions. Contrary to our hypothesis, Lombard speech exaggerates the surprisal category by increasing syllable duration and intensity range, suggesting that channel coding is expanded without compromising source coding.

5 Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. We thank Raphael Baier, Kirstin Kolmorgen, and Anna Konobelkina for annotation and technical support.

References

[1] LU, Y. and M. COOKE: *The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise*. *Speech Communication*, 51(12), pp. 1253 – 1262, 2009. doi:https://doi.org/10.1016/j.specom.2009.07.002.

[2] BRUMM, H. and S. A. ZOLLINGER: *The evolution of the lombard effect: 100 years of psychoacoustic research*. *Behaviour*, 148(11/13), pp. 1173–1198, 2011.

[3] AYLETT, M. and A. TURK: *Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei*. *The Journal of the Acoustical Society of America*, 119(5), pp. 3048–3058, 2006. doi:10.1121/1.2188331.

[4] CASTELLANOS, A., J.-M. BENEDÍ, and F. CASACUBERTA: *An analysis of general acoustic-phonetic features for spanish speech produced with the lombard effect*. *Speech Communication*, 20(1), pp. 23 – 35, 1996. doi:https://doi.org/10.1016/S0167-6393(96)00042-8.

[5] BORIL, H. and P. POLLÁK: *Design and collection of czech lombard speech database*. In *INTERSPEECH - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pp. 1577–1580. 2005.

[6] LU, Y. and M. COOKE: *Speech production modifications produced in the presence of low-pass and high-pass filtered noise*. *The Journal of the Acoustical Society of America*, 126(3), pp. 1495–1499, 2009. doi:10.1121/1.3179668.

[7] GARNIER, M. and N. HENRICH: *Speaking in noise: How does the lombard effect improve acoustic contrasts between speech and ambient noise?* *Computer Speech Language*, 28(2), pp. 580 – 597, 2014. doi:https://doi.org/10.1016/j.csl.2013.07.005.

[8] HANSEN, J. H. L., J. LEE, H. ALI, and J. N. SABA: *A speech perturbation strategy based on "lombard effect" for enhanced intelligibility for cochlear implant listeners*. *The Journal of the Acoustical Society of America*, 147(3), pp. 1418–1428, 2020. doi:10.1121/10.0000690.

[9] PATE, J. K. and S. GOLDWATER: *Talkers account for listener and channel characteristics to communicate efficiently*. *Journal of Memory and Language*, 78, pp. 1 – 17, 2015. doi:https://doi.org/10.1016/j.jml.2014.10.003.

[10] AYLETT, M. and A. TURK: *The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech*. *Language and Speech*, 47(1), pp. 31–56, 2004. doi:10.1177/00238309040470010201.

[11] JAEGER, T. F.: *Redundancy and reduction: speakers manage syntactic information density*. *Cognitive psychology*, 61(1), pp. 23–62, 2010. doi:10.1016/j.cogpsych.2010.02.002.

[12] HALE, J.: *Information-theoretical complexity metrics*. *Language and Linguistics Compass*, 10, 2016. doi:10.1111/lnc3.12196.

[13] GIBSON, E., R. FUTRELL, S. P. PIANTADOSI, I. DAUTRICHE, K. MAHOWALD, L. BERGEN, and R. LEVY: *How efficiency shapes human language*. *Trends in Cognitive Sciences*, 23(5), pp. 389 – 407, 2019. doi:https://doi.org/10.1016/j.tics.2019.02.003.

[14] HAZAN, V., O. TUOMAINEN, J. KIM, C. DAVIS, B. SHEFFIELD, and D. BRUNGART: *Clear speech adaptations in spontaneous speech produced by young and older adults. The Journal of the Acoustical Society of America*, 144(3), pp. 1331–1346, 2018.

[15] BRANDT, E., F. ZIMMERER, B. ANDREEVA, and B. MÖBIUS: *Mel-cepstral distortion of German vowels in different information density contexts*. In *Proceedings of Interspeech (Stockholm, Sweden)*, pp. 2993–2997. 2017.

[16] BARONI, M. and A. KILGARRIFF: *Large linguistically-processed web corpora for multiple languages*. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pp. 87–90. 2006.

[17] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer (version 6.1. 08) [computer program], retrieved may 1*. 2019.

[18] SCHIEL, F.: *Automatic Phonetic Transcription of Non-Prompted Speech*. In *Proc. of the ICPhS*, pp. 607–610. San Francisco, 1999.

[19] BATES, D., M. MÄCHLER, B. BOLKER, and S. WALKER: *Fitting linear mixed-effects models using lme4*. *Journal of Statistical Software*, 67(1), pp. 1–48, 2015. doi:10.18637/jss.v067.i01.

[20] BARR, D.: *Random effects structure for testing interactions in linear mixed-effects models. Frontiers in Psychology*, 4, p. 328, 2013. doi:10.3389/fpsyg.2013.00328.