ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Effect of noise type and level on focus related fundamental frequency changes

Martti Vainio[1], *Daniel Aalto*[1,2], *Antti Suni*[1], *Anja Arnhold*[1,3], *Tuomo Raitio*[2], *Henri Seijo*[2], *Juhani Järvikivi*[4], and *Paavo Alku*[2]

[1]Institute of Behavioural Sciences (SigMe Group), University of Helsinki, Finland
[2]Department of Signal Processing and Acoustics, Aalto University, Finland
[3]Department of Cognitive Linguistics, Goethe-University Frankfurt am Main, Germany
[4]Language Acquisition and Language Processing Lab, NTNU, Norway

martti.vainio@helsinki.fi

10.21437/Interspeech.2012-206

Abstract

Speech in noise, or Lombard speech, is characterized by increased intensity and higher fundamental frequency as well as lengthened segmental durations as speakers try to maintain a beneficial signal-to-noise ratio to fill both communicative and self-monitoring requirements. The phenomenon has been studied with regard to different noise types and different noise levels, as well as with respect to different communicative tasks (e.g., reading out loud vs. speaking to a real listener). However, there are no studies where the effect has been measured with different noises keeping the loudness levels equal. Here we study the Lombard effect with three different noise types at three levels with equal loudness while varying focus structure to elicit different pitch contours. The results show that people adapt their intonation contours depending on both noise level and type even when the noises are similar with respect to their perceived loudness. This points to a special role for pitch in Lombard speech.

Index Terms: Lombard speech, prosody, focus marking

1. Introduction

Speakers automatically raise their voice when forced to speak in environmental noise or when the normal feedback mechanism is disturbed. Raising one's voice consists of various physiological means that have different consequences on the phonetic realization of speech. Typically the speakers' f_0 is higher and the mode of vocal fold vibration is more pressed decreasing the slope of the glottal voice-source spectrum. The adaptation of speech to noise in order to increase the signal-to-noise ratio is called the Lombard effect or Lombard reflex to illustrate its involuntary nature [1]. With respect to communicative needs the reflex or effect is thought to involve both private and public feedback loops. That is, both speaker internal and speaker external features have been demonstrated to modulate the effect [2]. Although the Lombard effect was originally discovered with humans it has been attested in animals as well – thus it is partly a low-level biological phenomenon [3]. All mammals and birds are thought to be able of displaying the effect [4].

The knowledge regarding linguistic signaling in Lombard speech is fairly general in nature and not very much is known about how the reflex influences prosodic changes that are due to specific communicative needs such as signaling information structure. Generally the effect has been linked to signal amplitude (vocal intensity), whereas the voice fundamental frequency (f_0) has been regarded as a secondary feature whose raising has been seen to follow from increased intensity. The voice production in mammals and birds is due to similar processes involving vibrating membranes in either the mammalian larynx or the avian syrinx. Therefore it is difficult to assess whether the raising of f_0 is due to increased vocal intensity or vice versa. In any case, the raising of pitch has been attested in most studies of Lombard effect on humans as well as those conducted on birds [4].

There is indication that linguistic factors influence f_0 changes in Lombard speech. Patel and Schell reported that function words and content words behaved differently with regard to the increase in f_0 and duration [5]. Moreover, it was found in [6] that stressed and non-stressed words behaved differently with regard to the background noise in spontaneous speech.

The purpose of the study was to see whether speakers vary their prosodic means of marking focus as a function of both noise level and type. The analyzed utterances were replies to three types of questions designed to elicit either a broad focus or narrow focus on two different words (the first or last word) in simple three word utterances. The typical prosodic patterns for the three focus conditions are well-known for Finnish [7, 8], which allows us to compare Lombard speech to undisturbed speech in a controlled manner. The three types of noise were: babble noise, white noise, and a 1 kHz low-pass noise. The noises were scaled for equal loudness on three different levels corresponding to approximately 60, 70 and 80 dB(A) sound pressure levels.

With regard to prosody we were interested in the following questions: (1) How does noise affect f_0 contours in general? Apart from raised pitch, are the contours also expanded in f_0 range? (2) Are the changes affected by different types of noise and noise levels regardless of equal loudness? And (3) Do the linguistically motivated f_0 features remain the same as in speech without the Lombard effect?

2. Experiment

2.1. Materials and procedure

We recorded 21 speakers (11 female, mean age 26 years) producing utterances with different focus conditions in three types of noise with four noise levels. The participants were mostly students at either Aalto University or the University of Helsinki and none reported any hearing problems. The recordings were done in an anechoic chamber at the Aalto University using closed headphones for both noise playback and speech feedback for self-monitoring.

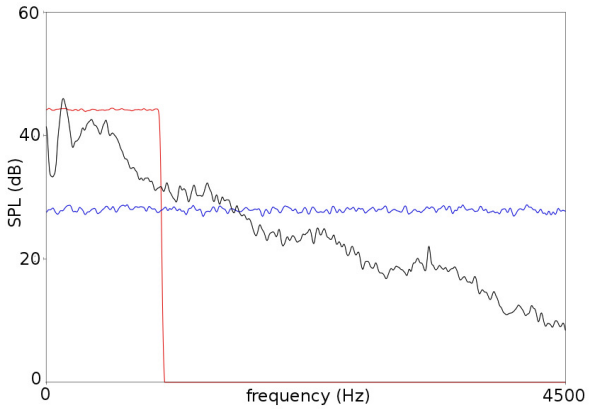

Figure 1: Representative spectra of the three different noise types used in the study; red = low-pass noise, black = babble noise, blue = white noise. The figure shows the spectra between 0 and 4.5 kHz.

Three focus conditions were elicited: *broad focus*, *narrow focus on subject* (the first word) and *narrow focus on object* (the last word). The f_0 contours for these conditions are well known and we expected them to remain unaltered [7, 8]. The utterances were of the form e.g., *Paavi tavaa suuraa* (The pope reads a sura), and long vowels [ɑ], [i], and [u] were used in the subjects and objects. With three focus conditions, as well as three vowels in two positions, 12 sentences were created. Each sentence was matched with a suitable question in order to elicit the correct focus condition. The resulting question-reply pairs were randomized and divided into nine separate lists with ten sets of twelve pairs. In case of narrow focus on either the subject or the object, the word was printed with bold letters. The participants read the sentences from paper sheets as if to reply to the question. The participants were instructed to speak clearly and each session was preceded by six trials. The noises were presented in a randomized order for at most 5 minutes at a time.

2.1.1. *Noises and system calibration*

Three different noise types were chosen for the study: babble noise, white noise, and low-pass filtered white noise. All noises were scaled to have equal loudness at three separate levels. Figure 1 shows smoothed spectra of the noise types calculated from approximately one minute segment of each signal. We chose white noise for maximal energetic masking, whereas the babble noise was chosen to reflect a situation with informational masking. The low-pass filtered noise extends to 1000 Hz and the cut-off point was chosen so that the noise would influence the vowels differently; i.e., both F1 and F2 would be masked in the case of [u], whereas only F1 would be masked in [i] and [ɑ].

The white noise was created with Matlab using 96 kHz sampling frequency and 24-bit quantization. The low-pass noise was then created by low-pass filtering the white noise using a 50-degree elliptic filter with 0.1 dB pass-band ripple and a 150 dB stop-band attenuation above 1 kHz. For the babble noise we used a noise taken from the NOISEX-92 data-base [9]. The original babble noise was resampled to have the same sampling frequency as the other noise signals. The original sampling frequency of the babble signal was 19980 Hz. The duration of the noises varied from 4 to 5 minutes.

The noises were scaled to have equal loudness values of 4.75, 9.5, and 19 sones corresponding roughly to 60, 70, and 80 dB(A) sound pressure levels. We used the ANSI S3.4-2007 [10, 11] standard for the noise scaling. The levels were calculated using 1/3 octave bands.

In order for the noise signals to be played at the desired loudness level for the participants their output levels had to be calibrated. For this we used an artificial head (Cortex Mk2) and high quality closed headphones (Sennheiser HD250 Linear II). The same headhones were used during the recording. The recording system was calibrated using an SPL-meter, loudspeaker and a microphone in an anechoic chamber. The recordings were done in the same anechoic chamber using a high-quality condenser microphone (AKG CK92) and a high-quality analogue to digital converter (Motu Traveler MkII).

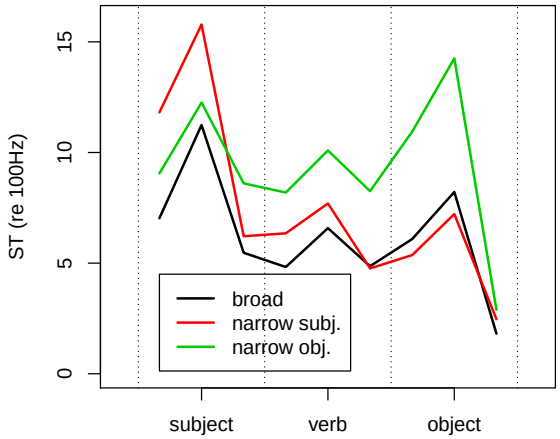

Figure 2: Average f_0 contours calculated from three f_0 points per each word for different focus conditions for all noise types and levels. Black = broad focus, red = narrow focus on subject, green = narrow focus on object.

3. Results

We analyzed the produced utterances with regard to f_0, *duration*, *voice source features*, *formants* and *intensity*. Only pitch related features are presented here. The pitch contours were analyzed in terms of three different points per word: the pitch maximum (peak) and the minima left and right of it (valleys). Thus, there are nine potential values for each utterance. Figure 2 shows averaged contours from the nine peak-valley points by focus condition. The contours clearly follow the typical shapes associated with different focus conditions in Finnish [7, 8]; i.e, the narrowly focused word has a higher peak and post-focal words have lower peaks but are not altogether deaccented. The verbs also have a rising-falling shape, but with a markedly lower magnitude [12]. For further analyses the f_0 values were transformed to semitones (re 100 Hz).

Figure 3 shows the averaged f_0 contours by noise level and noise type. It is clear from the figures that both level and type of noise have an effect on the overall f_0 level, although the difference between white and babble noise is relatively small.

The f_0 *expansion* was calculated from the nine f_0 values per sentence (also in semitones) by adding the absolute differences. This can be expressed in terms of an integral based on

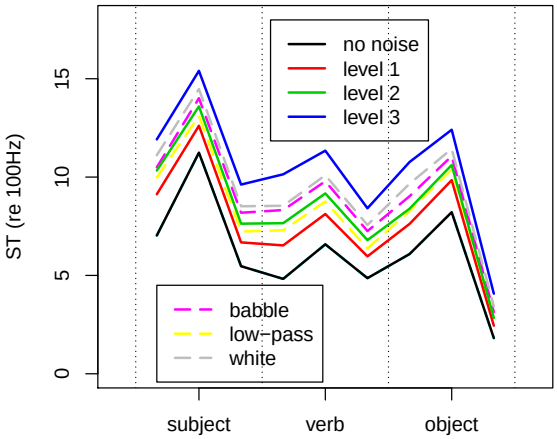

Figure 3: Average f_0 contours for different noise levels and types.

the Bounded Variation (BV) norm:

$$\text{Expansion}(f_0) = \int_{T_{beg}}^{T_{end}} \left| \frac{df_0(s)}{dt} \right| ds \qquad (1)$$

where $f_0(t)$ is the fundamental frequency at a given time point and T_{beg} and T_{end} are the beginning and end times of the utterance.

The BV norm captures the overall movement in the contour in a time-invariant manner. The use of the BV norm is inspired by the neurophysiology of the first processing steps of pitch in the brain stem. The f_0 is mainly coded in the periodicity of the auditory nerve signals and the periodotopic axis emerges in the central nucleus of the inferior colliculus [13]. There, the pitch frequencies are logarithmically arranged, and the BV norm (for short time intervals) can be interpreted as corresponding to the diameter of the f_0 activated neural population. With regard to the points of interest in the contour, the BV is simply calculated as the sum of the absolute differences between the points. Using semitones the calculation yields a value depicting the overall change in semitones during the utterance. By using the manually marked f_0 points we could avoid the problems caused by f_0 detection such as, e.g., octave errors. The calculation, however, does not account for the very low f_0 values related to creaky voice, which typically occurs at the end of utterances in Finnish. That is, there are a number of missing values due to non-modal voice in the data. The creaky voice, on the other hand, virtually disappears from the utterances in noise and should not present a problem in our analyses.

Figures 4 and 5 show the effect of noise level and type on both mean f_0 and mean f_0 expansion, respectively. Both were calculated from the nine points. The levels without noise were 6.63 ST for f_0 and 20.4 ST/utterance for the expansion.

Statistical analyses were done using linear mixed-models with participants and items as a crossed-random factor and focus type, noise level, noise type, and gender as fixed-effects predictors [14, 15]. Model selection was done using backward elimination and log likelihood ratio tests (function anova in R). Model comparison indicated that adding by-subject and by-random slopes for the fixed-predictors focus, noise level, and

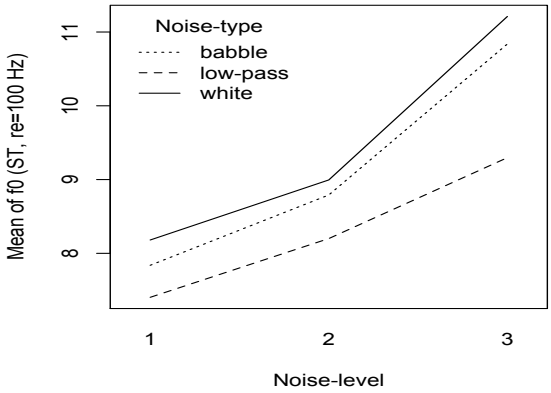

Figure 4: *Mean f_0 level vs. noise-level (no-noise = 6.63 ST).*

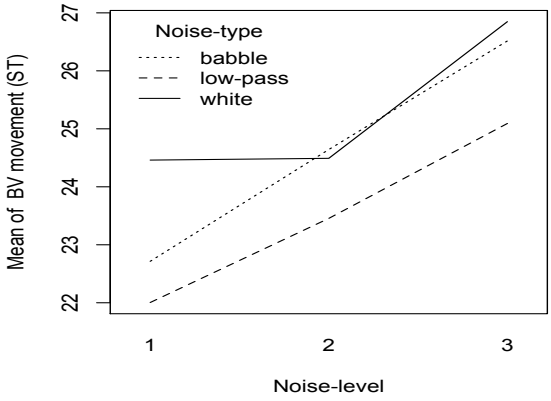

Figure 5: Mean BV movement vs. noise-level (no-noise = 20.4 ST).

noise type, significantly increased model fit. The final model for mean f_0 is depicted in Table 1 and for f_0 expansion in Table 2. In both tables the intercept stands for female speakers, broad focus condition and babble noise at 60 dB(A) (4.75 sones) level. The estimates are in semitones and noise levels 2 and 3 stand for 9.5, and 19 sones, respectively.

Table 1: Mixed-effects model results for mean f_0.

	Estimate	Std. Error	t-value
(Intercept)	13.33939	0.45707	29.185
focus N1	-1.01482	0.15516	-6.540
focus N2	0.69438	0.10047	6.912
low-pass	-0.43408	0.22300	-1.947
white	0.36448	0.20839	1.749
level2	0.95697	0.12598	7.596
level3	2.98460	0.29964	9.961
gender male	-11.33108	0.53456	-21.197
low-pass:level2	-0.20048	0.12605	-1.590
white:level2	-0.20695	0.12612	-1.641
low-pass:level3	-1.09177	0.12612	-8.657
white:level3	0.02307	0.12618	0.183

With regard to mean f_0 the noise levels differ significantly (t-values approximately 2 or greater). The low-pass noise has a significantly lower mean f_0 (t = -1.95). As can be expected from the results of previous studies, the different focus types are also different from each other: i.e, the f_0 is generally lower when the narrow focus occurs on the first word (N1) and higher when it occurs on the last word (N2). Also, the expected gender difference is highly significant with the males speaking almost an octave lower than the females. There is also a significant low-pass-noise:noiselevel3 interaction showing that the f_0 level is increased less in high level low-pass noise.

Table 2: Mixed-effects model results for f_0 expansion.

	Estimate	Std. Error	t-value
(Intercept)	18.6681	1.4037	13.300
focus N1	2.5975	0.6263	4.148
focus N2	6.3596	0.6593	9.646
low-pass n.	-0.7105	0.6267	-1.134
white n.	1.7450	0.4794	3.640
noiselevel 2	1.9829	0.5206	3.809
noiselevel 3	3.8448	0.5808	6.619
gender male	2.1418	1.7392	1.231
low-pass:level2	-0.4808	0.5677	-0.847
white:level2	-1.8979	0.5680	-3.341
low-pass:level3	-0.6886	0.5680	-1.212
white:level3	-1.3883	0.5683	-2.443

With regard to f_0 expansion the results show that the contours are significantly influenced by the focus type as well as noise levels. The low-pass noise, however, does not differ from babble noise, but the contours are again more expanded in white noise (t = 3.64). This is also shown in the white-noise:level interactions. There are no significant gender differences.

4. Conclusions

Many of the results presented here are as expected: the f_0 level rises as a function of noise level and the f_0 contours are more expanded in noise. In addition to the typical f_0 level increase, there is an exponential increase in f_0 expansion when the noise level increases. The expansion effect is similar to the f_0 level with regard to different noise types. However, we also found differences between noise types with the low-pass noise having a smaller influence on f_0 levels and white noise having a greater influence on the f_0 expansion, regardless of equal loudness.

The Lombard effect has traditionally been interpreted as a speaker's need to increase vocal intensity in the presence of noise in order to be heard over the noise by both herself and the recipient. In the current study we show that regardless of equal loudness – that is, in terms of equally perceived masking level – speakers still change their behavior with respect to f_0 depending on the type of noise.

The question arises, then, as to the reason why different types of noise have different effects on the production of pitch contours. One possible answer has to do with how pitch is perceived on the lowest level, i.e., on the basilar membrane, where the masking properties of the different noise types are directly comparable with mechanism of extracting f_0 from the speech signal. The auditory system is able to recognize speech patterns based on limited frequency band information. This is reflected in the current data: to get enough auditory feedback and to take into account the virtual listeners, it is not necessary to increase the vocal intensity/pitch as much as it is when a broader masker is present.

The fact that the noise types have different effects on f_0 points to a specific importance of pitch in auditory feedback. The results also suggest that the Lombard effect may have to do with higher f_0 just as much as it has to do with increased signal amplitude of intensity.

5. Acknowledgements

We would like to thank Heini Kallio for her diligent work in segmenting and labeling the data. The study has been supported by the Academy of Finland (proj. numbers 1128204, 128204, 125940, as well as the Lastu program), by Tekes (the Finnish Funding Agency for Technology and Innovation, proj. number 440054), and by the EU/FP7 project Simple4All.

6. References

[1] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 101-119, p. 25, 1911.

[2] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *Journal of Speech and Hearing Research*, vol. 14, no. 4, p. 677, 1971.

[3] J. Bradbury and S. Vehrencamp, "Principles of animal communication." *Behavioral Ecology*, vol. 12, pp. 283–286, 1998.

[4] H. Brumm and A. Zollinger, "The evolution of the lombard effect: 100 years of psychoacoustic research," *Behaviour, 148*, vol. 11, no. 13, pp. 1173–1198, 2011.

[5] R. Patel and K. Schell, "The influence of linguistic content on the Lombard effect," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 1, p. 209, 2008.

[6] C. Rivers and M. Rastatter, "The effects of multitalker and masker noise on fundamental frequency variability during spontaneous speech for children and adults." *Journal of Auditory Research*, 1985.

[7] M. Vainio and J. Järvikivi, "Focus in production: Tonal shape, intensity and word order," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. EL55–EL61, 2007. [Online]. Available: http://link.aip.org/link/?JAS/121/EL55/1

[8] A. Arnhold, "Multiple prosodic parameters signaling information structure: Parallel focus marking in finnish," in *Proceedings of the International Congress of Phonetic Sciences (ICPhS), Hong Kong*, 2011.

[9] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[10] ANSI, "ANSI S3. 4-2007. procedure for the computation of loudness of steady sounds," 2007.

[11] B. Glasberg and B. Moore, "Prediction of absolute thresholds and equal-loudness contours using a modified loudness model," *The Journal of the Acoustical Society of America*, vol. 120, p. 585, 2006.

[12] A. Arnhold, M. Vainio, A. Suni, and J. Järvikivi, "Intonation of Finnish verbs," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[13] G. Langner, "Periodicity coding in the auditory system," *Hearing Research*, vol. 60, no. 2, pp. 115–142, 1992.

[14] D. Bates and D. Sarkar, "Linear mixed-effects models using s4 classes," *See http://cran. r-project. org/web/packages/lme4/index. html*, 2006.

[15] R. Baayen, D. Davidson, and D. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of memory and language*, vol. 59, no. 4, pp. 390–412, 2008.