Effect of task on the intensity of speech in noisy conditions

Julián Villegas, Ian Wilson, and Jeremy Perkins (University of Aizu)

1 Introduction

Humans modify their vocal effort when speaking in noisy environments [1]. Compared to speech in quiet conditions, speech produced in the presence of noise is acoustically characterized by an increase in intensity [2], a decrease in speech rate [3], etc.

The Lombard effect (LE) seems to be caused by an involuntary response when the self-monitoring mechanism is hindered by an energetic masker [4]. This reflex has been investigated by Anderson [5] who found that the average latency for LE in subjects uttering 'ah' while listening to 3 dB intensity changes of a 100 Hz square wave was about 127 ms. Bauer et al. [6] found a mean latency of 157 ms, for speech level increases in speakers subjected to changes in the feedback level of their own voices. This effect has also been linked to an active response of speakers perceiving that their peers have difficulties understanding them [7].

Besides the relative weights of the self-monitoring challenges and the perceived communication difficulties, it is not clear how the task carried out by speakers may affect their Lombard production, or how the actual transition between speech in quiet and noisy conditions happens.

To examine this, we conducted an experiment in which participants engaged in 4 tasks varying on communication effort and goal existence. Concretely, we aim to answer the following questions: (i) Do the speech levels observed on quiet and noisy conditions vary depending on task? Specifically, do the levels vary depending on whether the task requires an active communication effort, and whether the task is goal oriented? (ii) Regardless of the possible level differences in quiet conditions, do the transitions between noise to silence (and vice versa) present differences across different tasks?, and related to this, (iii) Are there differences between transitions to noise and transitions to silence?

2 Experiment

A cohort of 18 paid Japanese students from the University of Aizu (5 females) participated in the experiment. They were on average 21 years old (SD = 1.86), and had normal hearing thresholds, as verified with a Maico MA25 audiometer. Participants were self-organized in friend pairs. Permission for performing this experiment was obtained following the University of Aizu ethics procedure.

The experiment was conducted in an anechoic chamber where the participants sat facing away from each other and separated by about 5.5 m. To record participants' voices, DPA 4088 headset microphones were used. Participants were also wearing Sennheiser HD 380 Pro headphones from which they could hear the background noise and their interlocutor's voices at ≈ 64 dB (A)(only in the interaction tasks), their own voices were never amplified. Speech of each pair was recorded with a resolution of 24 bits/48 kHz. Game boards, and written material were presented via iPads.

Participants were subjected to alternating periods of silence and Gaussian noise, 30 s long. Noise periods were presented at 84 ± 0.1 dB (A), as verified with a Brüel & Kjær 4153 ear simulator and a 2250-Light-G4 sound level meter.

We recorded and analyzed speech of participants engaged in: (i) free dialog, where participants were encouraged to talk as if they were in a familiar environment, (ii) playing a game (Battleship), in which participants were asked to win as many matches as possible, (iii) freely speaking (soliloquy) in which the participants were only instructed to speak whatever came to their minds, and (iv) text reading, where participants were instructed to read as many pages as possible without rushing and while speaking clearly. These four tasks were selected to be communication-oriented (i, ii) or not (iii, iv) and goal-oriented (ii, iv) or not (i, iii). In all cases, we stressed the importance of speaking during tasks.

In a given session, a pair of participants performed each task over a duration of 15 minutes. The actual task sorting for a given pair was determined using a balanced Latin-square design. Each session lasted between 60 and 90 minutes.

Practice trials were held before the actual experiment, so participants got used to the background

noise changes, as well as the tasks. In the case of the game, participants were asked to mark 'hits' and 'misses' of both participants (i.e., when the letter-number corresponded or not to an enemy position).

Instructions (in Japanese) as well as hearing screenings were provided in the control room. Then, participants were escorted to the anechoic chamber where they were assisted on wearing the mike and headphones. Finally, the experimenter left the anechoic chamber. The experimenter and the participants communicated using mikes and headphones, but the experimenter's voice was muted during recordings.

Reading material include Japanese translations of short stories such as "The North Wind and the Sun," and "The boy who Cried Wolf" [8]. For the game, a customized digital version of Battleship was used. This game is characterized by short and frequent interactions where participants need to utter letter names ($A..K$) and digits (1..10) to refer to positions on the board. To save time, 7 ship positions for both participants were previously determined by the experimenters; participants were able to mark their interactions using a virtual pencil. Finally, lists of topics for soliloquy and dialog were offered prior to the experiment, but subjects were prevented from bringing them to the experiment.

Recordings were high-pass filtered in Matlab, using a 1024 order FIR filter with a cutoff frequency of 80 Hz, and separated into mono files for each participant. Speech regions were automatically segmented using the method described in [9] into 'speech,' 'silence,' or 'other.' The segmenting algorithm classified regions with $< 20\%$ of the mean energy and longer than 100 ms as silence, other regions were classified as speech if they were longer than 260 ms. Finally, 'intensity' (here called level) was computed every 10 ms in Praat using a minimum f_0 of 100 and 75 Hz, for female and male speakers, respectively.

3 Results

Table 1 Mean speech levels for each task.

Task	quiet [dB (SPL)]	noisy [dB (SPL)]
Dialog	61.3	66.7
Game	62.6	68.3
Soliloquy	54.5	58.8
Text reading	56.0	59.3

The mean speech levels in quiet and noisy conditions are summarized in Table 1. The overall effect of goal and communication effort (CE) on speech level in quiet and noisy conditions was assessed by extracting speech segments from noisy and quiet periods between 12 s and 20 s after the background noise/silence onset. In the case of speech in noise, we subtracted from these period levels the mean level of each participant speech within the same task as a way to compensate for the disparate levels in quiet conditions.

In the absence of an energetic masker, participants produced louder level for tasks involving CE, as confirmed with a repeated measures ANOVA with *goal* (no goal, goal) and CE (no, yes) as factors [$F(1,17) = 71.14$, $p < .001$, $\eta^2_G = .296$]. An additional increase in speech level was observed when the task was goal-oriented [$F(1,17) = 10.49$, $p = .005$, $\eta^2_G = .016$]. This increase was significant for tasks not involving CE, as illustrated in the top panel of Fig. 1.

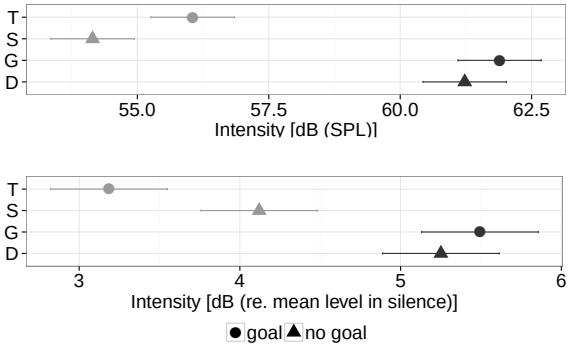

Fig. 1 Speech level differences between (T)ext reading, (S)oliloquy, (G)ame, and (D)ialog tasks in the absence of energetic masker (top) and in Lombard conditions (bottom). Error bars denote Fisher's Least Significant Difference (FLSD), disjoint bars indicate significant differences. Non-communicative tasks are shown in gray.

A similar ANOVA as before, revealed a significant effect of CE on speech levels in noisy conditions [$F(1,17) = 36.64$, $p < .001$, $\eta^2_G = .350$]. Participants spoke louder in noisy conditions when the task involved a communication effort. This increase in level was even higher if the task was also goal oriented. However, this trend is reversed when the task did not require communication effort. I.e., relative to their mean level in quiet conditions, participants produced significantly louder levels when reading as opposed to when freely speaking. This interaction between CE and *goal* was found to be significant

$[F(1,17) = 6.48, p = .021, \eta^2_G = .060]$, as well as the difference between text reading and soliloquy, as summarized in Fig. 1 (bottom panel). Note, however, that when considering the absolute levels in noisy conditions, whereas the effect of goal was not significant ($p = .060$), the effect of communication was still significant $[F(1,17) = 124, p < .001, \eta^2_G = .363]$. This suggests that in tasks with no communication, speakers were increasing their speech intensity to the same level regardless of the task goal.

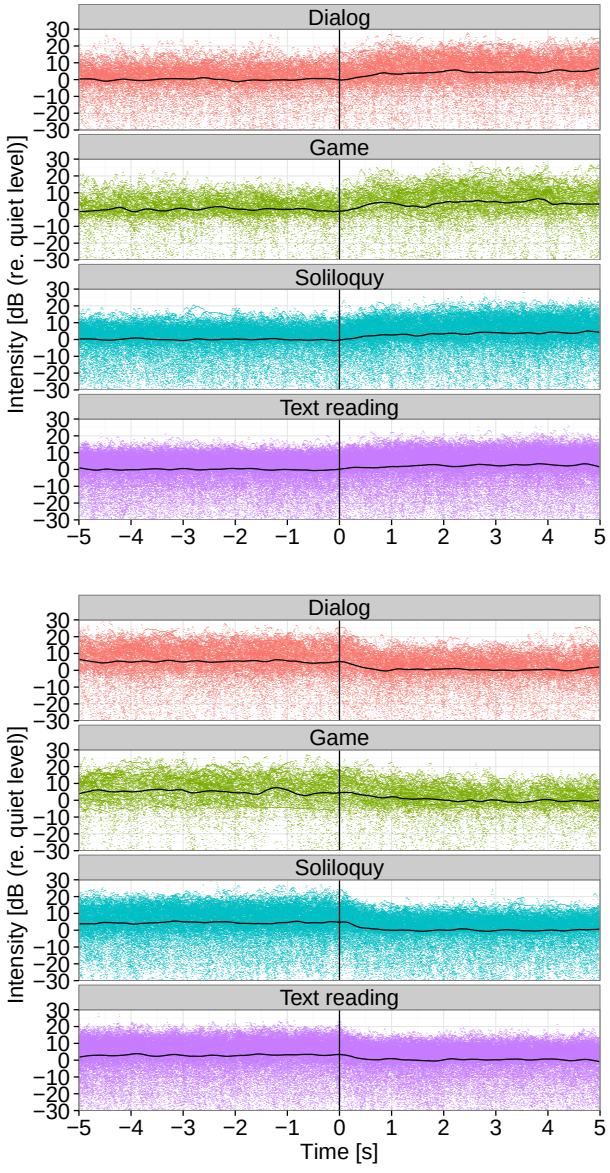

Fig. 2 Measured levels plotted along with a LOESS smoothing across all speakers for the four different tasks on the silence to noise transitions (top) and noise to silence transitions (bottom).

We extracted the silence-noise and noise-silence transitions (±5 s around the onset of the noise and silence, respectively) and for each group, we aggregated all the transitions per subject and task. Since we are interested in finding possible contour differences between tasks and their overall levels could hide such differences, the speech level of each speaker in silent periods preceding the noise onset on a given task was used as reference to compute the speech level in the transitions. Fig. 2 shows the actual level points for each task along with a local linear regression fitting contour.

Data corresponding to 1 s before the background change to 5 s after, were used for fitting a smoothing cubic spline ANOVA model as implemented in [11]. Smoothing parameters were selected by a generalized cross-validation method using the default smoothing factor $\alpha = 1.4$.

In our model, the level of speech is explained by the additive model of factors *task* and *time*. Their interaction was excluded after finding it not significant with a Kullback-Leibler projection (KL-ratio = .114 for silence-noise and KL-ratio = .093 for noise-silence transitions). The resulting spline shapes comprise the intercept (an additive constant), the main effect (the spline that best fits all the data regardless of task), and the task effect (an additive constant). Differences were considered significant if the Bayesian 95% confidence intervals around contours were disjoint.

The main effect of the fitted model for silence-noise transitions indicates that, on average, speech level raised over 3 dB for the first time at 900 ms after the noise onset, it continued to raise until flattening out at about 2000 ms. A further inspection of each task effect, revealed a non-significant effect of soliloquy, i.e., the difference between the effect of soliloquy and the main effect was −0.05 dB. Furthermore, effects of dialog and game were 0.48 dB and 0.36 dB, respectively, and not significantly different from each other. The text reading effect was found to be −0.78 dB.

On the noise-silence transitions, the main effect shows that speech level lowered by 3 dB for the first time at about 600 ms, reaching average speech levels on quiet conditions at about 1400 ms. The effect of task in these transitions was smaller than in the silence-noise transitions, but it was complex: whereas soliloquy and game effects were not significantly different from the main effect (0.03 dB and −0.03 dB, respectively), the text reading (−0.26 dB)

and dialog (0.21 dB) tasks were significantly lower and higher, respectively.

4 Discussion

The interpretation of these results should be done with caution since they are based on an automatic classification of speech and silence regions.

According to the results, it is more likely that speakers maintain a given speech level (across noisy and silent periods) while reading than while talking without a partner. Text reading had the least variation between noisy and quiet periods. We included it as a goal-oriented, non-communicative task, but these results could indicate that this task was rather mechanical, and since the goal-oriented nature of it was unrelated to communication, this may explain the lack of variation in the intensity between noisy and quiet conditions.

The transition analyses indicate that the increase (or decrease) in speech level continues beyond the previously reported reflex periods [5, 6], clearly indicating that speaking louder (or softer) is not simply a reflex, but is actively done by the speaker. They also indicate that noise-to-silence transitions were faster than silence-to-noise transitions.

5 Conclusions

The results presented here suggest that speech levels in quiet and noisy conditions vary depending on the task performed by the speakers, with communicative tasks yielding the highest values on both conditions. These results also suggest that speakers were faster in achieving mean levels in transitions from noise to silence than they were in the silence to noise transitions. The effect of task was also greater in the latter transitions, i.e., regardless of task, speakers returned to quiet levels in a more similar way than they did in the opposite direction.

According to these results, communication effort and, to a lesser extent, goal may play a limited role in the modulation of speech levels and that speakers may compromise vocal effort and self-monitoring capabilities or communication effectiveness depending on the task.

参考文献

[1] H. Brumm and S. A. Zollinger, "The evolution of the Lombard effect: 100 years of psychoacoustic research," *Behaviour*, vol. 148, pp. 1173–1198, 2011.

[2] E. Lombard, "Le signe d'élévation de la voix (the evidence of voice raise)," *Annales des maladies de l'oreille et du larynx*, vol. 37, pp. 101–119, 1911, in French.

[3] M. Cooke, C. Mayo, and J. Villegas, "The contribution of durational and spectral changes to the Lombard speech intelligibility benefit," *J. Acoust. Soc. Am.*, vol. 135, no. 2, pp. 874–883, Feb 2014.

[4] A. S. Therrien, J. Lyons, and R. Balasubramaniam, "Sensory Attenuation of Self-Produced Feedback: The Lombard Effect Revisited," *PLoS ONE*, vol. 7, no. 11, p. e49370, 2012.

[5] K. R. Anderson Foery, "Triggering the Lombard Effect: Examining Automatic Thresholds," Master's thesis, University of Colorado at Boulder, 2008.

[6] J. J. Bauer, J. Mittal, C. R. Larson, and T. C. Hain, "Vocal responses to unanticipated perturbations in voice loudness feedback: An automatic mechanism for stabilizing voice amplitude," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2363–2371, 2006.

[7] M. Garnier, M. Dohen, H. Loevenbruck, P. Welby, and L. Bailly, "The Lombard Effect: a physiological reflex or a controlled intelligibility enhancement?" in *7th Int. Sem. on Speech Production*. HAL - CCSD, 2006.

[8] D. Deterding, "The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation," *J. of the Int. Phonetic Assoc.*, vol. 36, pp. 187–196, 2006.

[9] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell System Technical J.*, vol. 54, no. 2, pp. 297–315, 1975.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2015, available [Mar. 2015] from www.praat.org.

[11] C. Gu, "Smoothing Spline ANOVA Models: R Package gss," *J. of Statistical Software*, vol. 58, no. 5, pp. 1–25, 2014.