Effects of noise and L2 on the timing of turn taking in conversation

A JOSEFINE MUNCH SØRENSEN[1,*], MICHAL FERECZKOWSKI[1,2], AND EWEN N MACDONALD[1]

[1] *Department of Health Technology, Technical University of Denmark (DTU), DK-2800 Kgs. Lyngby, Denmark*

[2] *Institute of Clinical Research, University of Southern Denmark, DK-5230 Odense M, Denmark*

Previous studies of floor-transfer offsets (FTO), the offset from when one talker stops talking to the next one starts, suggest that normal conversation requires interlocutors to predict when each other will finish their turn. We hypothesized that increasing the difficulty of holding a conversation by adding noise and/or speaking in a second language (L2) would result in longer FTOs. Conversations from 20 pairs of normal-hearing (NH), native-Danish talkers were elicited using the Diapix task in four conditions consisting of combinations of language (Danish vs. English) and noise background (quiet vs. ICRA 7 noise presented at 70 dBA). Overall, participants took longer to complete the task in both noise and in L2 indicating that both factors reduced communication efficiency. In contrast to our predictions, in the presence of noise, the median of the FTO distribution decreased by approximately 40 ms and the standard deviation decreased by approximately 60 ms. However, the average median duration of utterances increased by 40% in noise. These findings are consistent with talkers holding their turn for longer, which may allow more time for their own speech planning. Overall, the results suggest that talkers may prioritise maintaining social norms for turn-taking fluency when communicating in difficult environments.

INTRODUCTION

When talkers take turns (i.e., there is a transfer of who has the floor), the acoustic signals produced by each talker may partially overlap or be separated by a silent gap. The length of this interval is termed the floor-transfer offset (FTO) with a negative value indicating an overlap and a positive indicating a gap. In previous studies of conversational interaction, the FTOs are shorter than the latencies of speech planning and articulation, suggesting that in addition to conducting speech understanding and speech planning in parallel, interlocutors must also predict when their partner will end their turn (Levinson and Torreira, 2015; Stivers *et al.*, 2009). In difficult communication conditions (e.g., in the presence of noise), speech understanding may require more cognitive resources, resulting in fewer resources available for speech

*Corresponding author: ajso@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

planning and reduced saliency of the acoustic cues used to predict turn ends (for a discussion of these cues see Gravano and Hirschberg, 2011). As a consequence, we hypothesized that speech planning may be delayed and prediction precision may be reduced, which would be observed by a shift to the right and a broadening of the FTO distribution. To test these hypotheses, conversations were recorded in conditions that varied in the degree of expected communication difficulty. Conversations were recorded both in the absence and presence of background noise, with talkers speaking both in their native language (Danish) and in a second language (English).

METHODS

Participants

40 young normal-hearing (NH) native Danish talkers ($\mu_{\text{age}} = 26$ years, $\sigma^2 = 3.7$ years, 12 women) participated in acquainted pairs (four mixed-gender pairs). All participants had hearing threshold levels below 20 dB HL between 125 Hz and 8 kHz, self-reported as being "comfortable" in English, and had all participated in at least one university-level class taught in English. All participants provided informed consent and the experiment was approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). The participants were compensated for their time.

Setup and Recordings

The participants were seated in separate listening booths and wore Sennheiser HD650 open headphones and head-worn Shure WH20 microphones. The microphone levels were calibrated such that the resulting presentation levels over headphones were the same as if their interlocutor was one meter away from them in the same room. All calibrations were done in dBA. The noise presented in the experiment was a 6-talker speech-shaped noise (ICRA 7, Dreschler *et al.* (2001)) calibrated to an average presentation level of 70 dBA. In the headphones, the participants heard a linear mix of themselves, their interlocutor, and the background noise (in noise conditions).

Task and Procedure

The participant pairs elicited dialogue by conducting the Diapix UK task (Baker and Hazan, 2011), a spot-the-difference task in which participants are given almost identical cartoon pictures, and they have to work together to find the differences between them. To familiarize participants with the task, they conducted a Diapix task using pictures from the original Diapix corpus Van Engen *et al.* (2010) facing each other outside the audiometric booths under the experimenter's supervision. Following this, they moved to the two separate booths and conducted a second Diapix task in background noise. The test session consisted of finding 10 differences between Diapix UK pairs in three repetitions of four conditions consisting of the combinations of conversing in either their first (L1, Danish) or second language (L2, English) in quiet or noise. The order of the conditions was randomized within each replicate. After each

replicate the participants had a break. All the recordings for which we have received consent have been made publicly available (Sørensen *et al.*, 2018).

Analysis of Recordings

Voice activity detection was performed on the individual microphone tracks to get binary speech activity arrays. For each conversation, the binary speech arrays from the two talkers were fed into a classification algorithm developed by the author. The algorithm categorized the conversations into utterances (speech tokens by each individual separated by silence of less than 180 ms), gaps (joint silence following a floor transfer), overlaps-between (joint speech during floor transfer), overlaps-within (joint speech during the utterance of one talker that does not result in a floor transfer), and pauses (joint silence not followed by a floor transfer), see Fig. 1. For analysing the effects of noise, second language and replicate on various measures, mixed-effects regression models were fitted to the variables of interest using the *lme4* package in *R*. ANalysis Of VAriance (ANOVA) tables were provided with Satterthwaite approximated denominator degrees-of-freedom (df) corrected F-tests for the fixed effects. The *lsmeans* function from the *lmerTest* package was used to compute pairwise comparisons of least-squares means of the significant effects using the Satterthwaite approximated df.

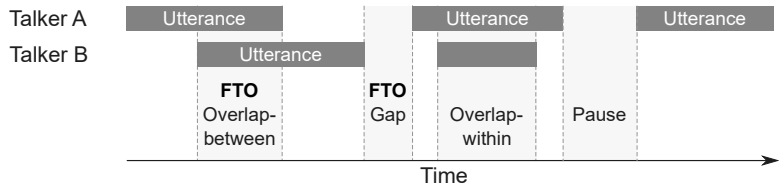

Fig. 1: Illustration of the classification of gaps, overlaps-within, overlaps-between, pauses, and utterances during conversations between Talker A and B. There are two floor-transfer offsets (FTOs): the overlap-between and gap.

RESULTS

Speech levels at a one-meter distance from each talker were estimated by computing the root mean square (RMS) level of all speech units, excluding pauses, and are plotted in Fig. 2, right panel. On average, talkers spoke 8.9 dBA louder in background noise than they did in quiet, resulting in an average SNR of -2.5 dB. There was a main effect of background [$F(1,39) = 1123.6, p <$ 2e-16], and of replicate [$F(1,39) = 3.91, p < 0.0283$]. A multiple comparison post-hoc analysis revealed that there was only a significant difference in level between replicate 1 and 3 [$t(39) = 2.648, p < 0.0116$], where the talkers spoke significantly softer by on average 0.44 dBA in the third replicate.

The task-completion time, i.e. time it took each pair to find 10 differences between the Diapix, was measured as a proxy for communication efficiency. The left panel

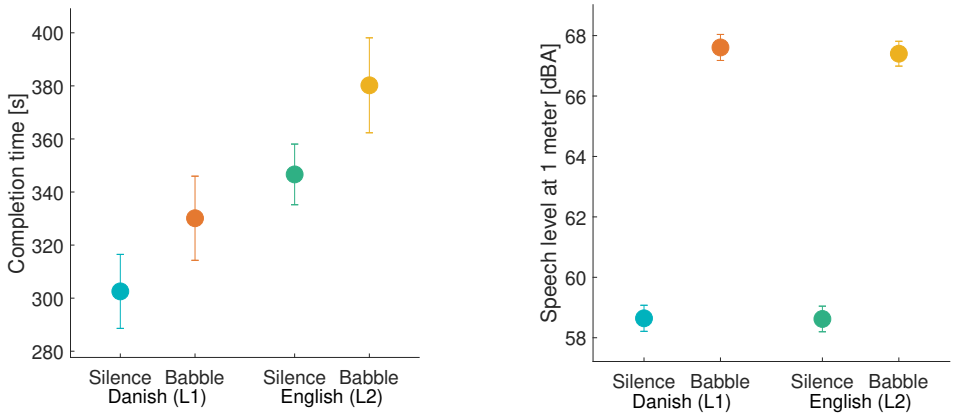

Fig. 2: Average completion time (left panel) and speech level (right panel) produced in the four combinations of language and noise. The bars indicate standard error.

of Fig. 2 shows the average completion time in the four conditions. There was a statistically significant training effect, i.e. the average completion time decreased with replicate [$F(2,216) = 23.1, p <$ 8.35e-10]. A pairwise comparison revealed a significant decrease in completion time between first and second replicate [$t(216) = 4.68, p <$ 5e-6], but only a borderline significant decrease between second and third replicate [$t(216) = 1.92, p < 0.0563$]. The completion time in noise compared to quiet increased significantly by 31 s [$F(1,216) = 12.3, p <$ 5.46e-4]. Similarly, participants took on average 47 s longer to complete the task in L2 compared to L1 [$F(1,216) = 29.2, p <$ 1.71e-7].

In all conversations, the articulation rates of the individual talkers were computed using the Praat script presented in Jong and Wempe (2009) with default parameter settings. The rate is measured as the number of syllables in the portions of the recording containing speech (i.e., periods of silence are excluded from the analysis). The average articulation rates are depicted in Fig. 3, left panel. The decrease in articulation rates in L2 compared to L1 was statistically significant [$F(1,438) = 680.02, p <$ 2.2e-16], and articulation rates increased significantly in noise [$F(1,438) = 60.3, p <$ 5.95e-14].

The average rate of overlaps-within (Fig. 3, right panel) increased both in L2 [$F(1,216) = 10.9, p < 0.00114$] and in noise [$F(1,216) = 77.7, p <$ 4.09e-16], and decreased with replicate: [$F(2,216) = 3.35, p < 0.037$]. A pairwise comparison revealed a significant decrease in rate between first and second replicate: [$t(216) = 2.42, p < 0.0165$], but not between second and third replicate [$t(216) = -0.408, p < 0.684$].

The overall hypothesis was that with increased processing demands, we would see

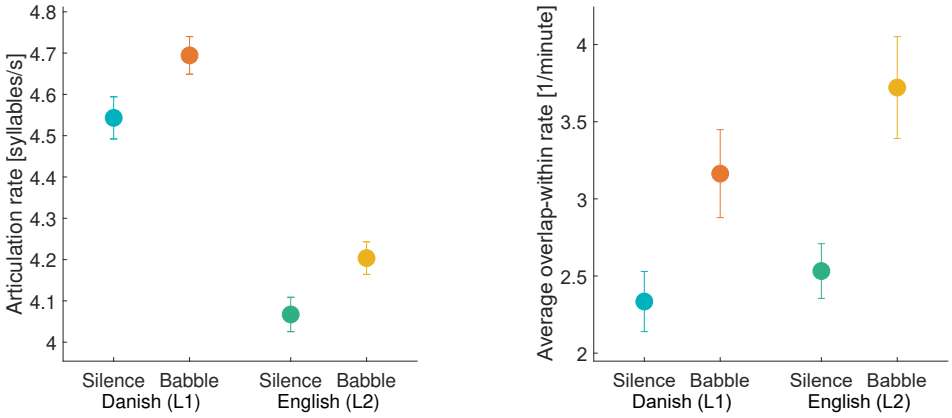

Fig. 3: Average articulation rates (left panel) and rate of occurrence of overlaps-within (i.e., turns from one talker that occur completely within a turn of the other talker) (right panel) produced in the four combinations of language and noise. The bars indicate standard error.

a delay and more variability in the timing of people's turn-taking. As a measure of centrality of the distribution, the median was used rather than the mean as FTO distributions are slightly positively skewed. There was a statistically significant main

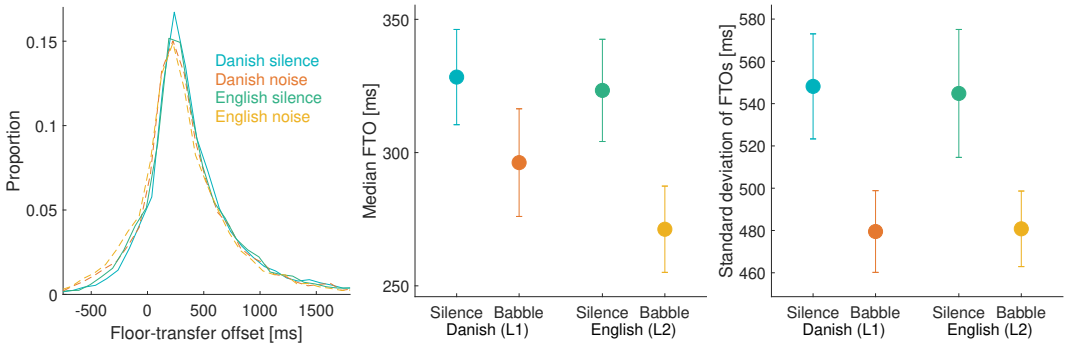

Fig. 4: Normalized distributions (left panel) of floor-transfer offsets (FTOs) along with the median (middle panel) and standard deviation (right panel) for the four combinations of language and noise. The bars indicate standard error.

effect of background [$F(1,218) = 46.6, p <$ 8.67e-11] and language [$F(1,218) = 5.91, p < 0.0159$] on the median FTO. Opposite to our hypothesis, the median decreased in the presence of background noise (by 40 ms, on average) and in L2 (by 15 ms, on average). The standard deviation averaged across pairs is plotted in Fig. 4. The analysis was performed on log-scaled standard deviations as the residuals were

not normally distributed, and there was a statistically significant effect of background [$F(1,219) = 21.5, p <$ 6.05e-6]. On average, the standard deviation decreased by 60 ms in background noise. In noise, interlocutors lengthened their turns substantially as

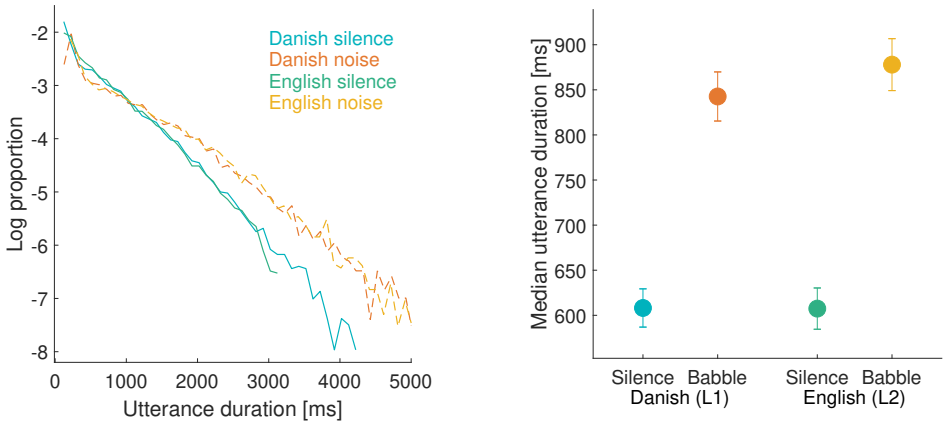

Fig. 5: Normalized distributions (left panel) and median (right panel) of the duration of utterances for the four combinations of task and noise. The bars indicate standard error. Note that the density in the left panel has been log-transformed to more easily compare the slopes.

is shown by a significant increase of about 42% of the median utterance duration in noise: [$F(1,219) = 738.71, p <$ 2.2e-16], depicted in Fig. 5, right panel. The analysis was performed on log-transformed utterance durations. The shallower slope of the pooled utterance durations across pairs as seen in Fig. 5, left panel, indicates a general lengthening of utterances in noise.

DISCUSSION

We hypothesised that communicating in noise and L2 would lead to an FTO distribution that was shifted to the right and broader due to an increase in cognitive processes used to understand and produce speech. We observed the opposite. In both noise and L2 the median FTO decreased, and the standard deviation decreased in noise. However, the changes were very small: the median decreased by about 40 ms in noise and 15 ms in L2, and the standard deviation decreased by about 60 ms in noise. Moreover, the overall shape of the distributions were similar across conditions and are similar to what has been found in other studies (e.g., Levinson and Torreira, 2015; Stivers *et al.*, 2009). This suggests that this turn-taking behaviour is important and that there may be a universal pattern in how we take turns that is more important to maintain than other properties of the speech we produce.

If maintaining rapid turn switches is important for social interaction, people may initiate their turns without having fully planned them yet. For example, they could

start with a filler-word, and then figure out what to say as they proceed with their sentence, leading to longer utterances. We saw, indeed, a substantial lengthening of the participant's turns in background noise. We also observed faster articulation rates in noise. Since speech planning is 3-4 times faster than articulation (Wheeldon and Levelt, 1995), longer utterances give the talker more time to continue speech planning.

In L2, participants decreased their rate of articulation and took longer to complete the task than in L1. However, unlike noise, L2 did not affect utterance duration or speech level. This is somewhat surprising as conversing in L2 should increase the processing loads for both speech perception and production, whereas noise should only affect the processing load for speech perception.

In Sørensen *et al.* (2020), when talking to a hearing-impaired (HI) interlocutor, the median FTO of both NH and HI increased in noise, and the FTO distributions became broader. While increased utterance durations were also found in that study, as well as other adaptations such as talking slower and overlapping less, their strategies may not have been enough to overcome the increased communication difficulty imposed by the hearing loss, leading to longer and more variable response times. In this study, however, the adaptive behaviour of the participants may have been sufficient to maintain "normal" turn switching behaviour.

Different, and sometimes opposite adaptive behaviours in noise have been observed in other conversation studies. For example, while Beechey *et al.* (2018); Sørensen *et al.* (2020) and the present study observed increases in utterance duration in noise, Hadley *et al.* (2019) observed the opposite. While there are several methodological differences across these studies, such as differences in the task or whether the average background noise levels were switched regularly or held fixed for several minutes, it is not yet clear which factors are responsible.

An increase in articulation rate, the amount of overlaps-within and the small decrease in median and standard deviation of FTOs in the noise condition may be an indication that interlocutors in the present study tried to move their operating point towards more rapid interaction. The increase in completion time, however, suggests that those changes did not maintain communication efficiency.

SUMMARY

Normal-hearing participants took longer to solve the Diapix task both in L2 vs. L1 and in noise vs. quiet. The median and standard deviation of FTOs decreased slightly in the presence of noise. An increase in the average utterance duration of about 40% in noise indicated that participants held their turn longer, allowing more time for their own speech planning. This suggests that interlocutors prioritize maintaining turn-taking fluency when adapting their behaviour in challenging acoustic environments.

ACKNOWLEDGEMENTS

A.J.M.S. and a portion of this study was supported by the William Demant Foundation (16-3968).

REFERENCES

Baker, R., and Hazan, V. (**2011**). "DiapixUK: Task Materials for the Elicitation of Multiple Spontaneous Speech Dialogs," Behav. Res. Methods, 43(**3**), 761–70. doi: 10.3758/s13428-011-0075-y

Beechey, T., Buchholz, J.M., and Keidser, G. (**2018**). "Measuring communication difficulty through effortful speech production during conversation," Speech Commun., **100**, 18-29. doi: 10.1016/j.specom.2018.04.007

de Jong, N. and Wempe, T. (**2009**). "Praat script to detect syllable nuclei and measure speech rate automatically,"Behav. Res. Methods, **41**, 385-90. doi: 10.3758/BRM.41.2.385

Dreschler, W., Verschuure, H., Ludvigsen, C. and Westermann, S. (**2001**). "ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment," Int. J. Audiol., 40(**3**), 148–57. doi: 0.3109/00206090109073110

Gravano, A., and Hirschberg, J. (**2011**). "Turn-taking cues in task-oriented dialogue", Comput. Speech Lang., **25**, 601–634. doi: 10.1016/j.csl.2010.10.003

Hadley, L. V., Brimijoin, W. O., and Whitmer, W. M. (**2019**). "Speech, movement, and gaze behaviours during dyadic conversation in noise," Sci. Rep., **9**, 10451. doi: 10.1038/s41598-019-46416-0

Levinson, S. C., and Torreira, F.(**2015**). "Timing in turn-taking and its implications for processing models of language," Front. Psychol. **6**, 731. doi: 10.1038/s41598-019-46416-0

Sørensen, A. J. M, Fereczkowski, M., and MacDonald, E. N. (**2018**). "Task dialog by native-Danish talkers in Danish and English in both quiet and noise," Dataset. doi: 10.5281/zenodo.1204951

Sørensen, A. J. M., Lunner, T., and MacDonald, E. N (**2020**). "Timing of turn taking between normal-hearing and hearing-impaired interlocutors," Proc. ISAAR, **7**, 37-44.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J.P., Yoon, K., Levinson, S.C. (**2009**). "Universals and cultural variation in turn-taking in conversation," Proceedings of the National Academy of Sciences Jun 2009, **106**(26), 10587-10592. doi: 10.1073/pnas.0903616106

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., and Bradlow, A. R. (**2010**). "The Wildcat Corpus of Native-and Foreign-accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles," Lang. Speech, **53**(4), 510–540. doi: 10.1177/0023830910372495

Wheeldon, L. R., and Levelt, W. J. M. (**1995**). "Monitoring the time-course of phonological encoding," J. Mem. Lang., 34, 311–334. doi: 10.1006/jmla.1995.1014