Measuring communication difficulty with eye-gaze behavior when speaking and listening

Susan Aliakabryhosseinabadi[1,2], Gitte Keidser[1,3], Tobias May[2], Torsten Dau[2], Martin A. Skoglund[1,4], Sergi Rotger-Griful[1]

[1] *Eriksholm Research Centre, Oticon A/S, Snekkersten, Denmark*

[2] *Hearing system section, Department of Health Technology, Technical University of Denmark, Ørsted Plads, Building 352, DK-2800 Kgs. Lyngby, Denmark*

[3]*Department of Behavioral Sciences and Learning, Linnaeus Centre HEAD, Linköping University, Linkoping, Sweden*

[4]*Division of Automatic Control, Department of Electrical Engineering, Linköping University, Linköping, Sweden*

Introduction

Eye-gaze behavior, referring to where a person looks at, plays an important role in social interactions, including emotional recognition and social attention and is sometimes studied in people suffering from a disability, like autism spectrum disorder or people with limited motor abilities [1], [2]. In the field of cognitive hearing science, eye-gaze behavior has been considered as a control signal in hearing aids [3], [4] and to understand cognitive process allocation in listening tasks [5]–[7]. For example, Sabic et al. found that gaze patterns in audiovisual stimuli are sensitive to a simulated hearing impairment, noise level and number of talkers when *following a conversation* [8]. More recently, there has been a growing interest in studying eye-gaze behavior during *interactive conversations* to examine how eye-gaze behavior is affected by hearing impairment and background noise [7], [9], [10].

Hadley et al. studied the effect of background noise in dyadic conversations between two hearing-impaired (HI) participants [10]. They found that at lower noise levels, listener's eye-gaze focused more on the *eye region* of the talker while at higher noise levels listener's eye-gaze focused more on the *mouth region* of the talker. Lu et al. studied the effect of hearing loss and background noise level on triadic conversations: two normal hearing (NH) confederates conversing with a NH or HI participant [9]. They found that the eye-gaze of HI participants focused less on the confederate that was actively talking when compared to the NH participants. Furthermore, they also found that only the NH participants accurately looked at the active talker (i.e., talker that was speaking) at high noise levels.

In this publication, we studied eye-gaze behavior in dyadic conversations between NH participants under different noise levels and hearing status. To engage participants in a dialogue, we used the Diapix task, in which participants need to find differences between two nearly identical pictures by talking to each other [11]. By means of speech analysis, we investigated eye-gaze behavior when participants were *passively listening* to their interlocutors and when participants were *actively speaking*. We aimed to examine how eye-gaze fixation and saccade would change in more challenging communication conditions (e.g., in higher background noise) to solve the task, not only across entire communication time but also within speaking and listening time separately.

Methods

Participants

Twenty-four elder (63.2±6.4 years) NH participants (13 females and 11 males) were recruited and divided into 12 pairs for a dyadic conversation. Participants were native Danish speakers and had age-adjusted normal hearing thresholds according to ISO-7029 with an average threshold of 26 dB HL, ranging from 20 to 40 dB HL. The study was approved by the Science-Ethics Committee for the Capital Region of Denmark (reference number H-16036391) and all participants signed a consent form before participating in the test.

Test setup

The face-to-face conversation was organized in dyads. The task of the participants was to spot the differences between two Diapix pictures (nearly identical pictures), copied from the original Diapix corpus [11] with Danish signage and exclamations introduced. Participants had a maximum of four minutes to find up to 12 differences. Should they find all differences before four minutes, the trial stopped, and the completion time was noted.

The experiment took place at Eriksholm Research Centre in a lab equipped with eight equally distanced loudspeakers in a horizontal ring and eight Vicon Vero motion capturing cameras. Participants were seated by a table at 1.5 m from each other and were equipped with a close-mouth microphone (DPA 4488, Germany) to record their speech, and Tobii Pro3 glasses to record their eye-gaze. Two sets of reflector markers, placed on both sides of the Tobii glasses, were used to track participants' head movements by the Vicon cameras.

Participants carried out the test under different conditions with respect to hearing status and background noise level. *Hearing status* was manipulated by asking both participants to wear a pair of earplugs (Alpine, MusicDafe Pro) thus simulating a mild (25 dB, on average), simulated hearing loss (SHL). *Background noise* level was manipulated with normal hearing condition by presenting babble noise from the loudspeaker array at 60 dBA (NH-N60) and 70dBA (NH-N70). Participants completed the test twice under each of the following conditions:

- NH-N0: This was the reference condition without earplugs and no noise.
- SHL-N0: With earplugs and no noise.
- NH-N60: Without earplugs and noise at 60 dBA.
- NH-N70: Without earplugs and noise at 70 dBA.

Data analysis

Speech analysis

The audio tracks from each talker were processed using a Voice Activity Detection (VAD) technique to extract speech/no-speech segments [12]. The speech signals were divided into segments of 5 ms duration and were labeled as speech (VAD = 1) if the RMS value exceeded a threshold, else they were labeled as non-speech (VAD = 0). The speech threshold was defined individually for each audio track and a subset of all audio tracks were manually checked to ensure proper labeling.

Speaking and listening time

Speaking time was defined as a speech segment (VAD = 1) or sequence of speech segments of the same speaker separated by a small pause (lower than 300 ms). *Speaking times* of less than 500 ms duration were excluded from the analysis as they were usually attributed to artifacts like coughing. Likewise, isolated *speech segments* of less than 100 ms were re-labeled as non-speech (VAD = 0). Note that speaking time could include overlap periods (i.e., both talkers speaking at the same time). Speaking time for one participant constituted the corresponding *listening time* for the other participant.

Eye-gaze analysis: saccades and fixations

A *saccade* refers to a fast eye movement occurring by switching the gaze between two objects, or regions-of-interest (ROIs). *Saccades* were computed from the angular velocity of eye-gaze in the vertical plane obtained from data collected by Tobii Pro 3 glasses. If the number of missing values in each trial was less than 30% of all samples, it was considered for further analysis. Before computing the derivative of the signal, a second-order pass-band Butterworth filter removed high frequency noise components of the signal. We then assumed a saccade happened if: (1) the vertical gaze velocity was over a threshold, individually selected for each participant; and (2) the saccade duration was in the range of 100 to 300 ms [4]. From this analysis, we obtained the *number of saccades*.

We assumed two main ROIs in the scene: the Diapix picture (placed on the table in front of the participants) and the interlocutor. Furthermore, we assumed that when no saccade happened, the eye-gaze remained *fixated* to either the picture or the interlocutor. To know whether the fixation was on one or the other, we combined data from the Tobii Pro 3 glasses and the Vicon motion capture. To that end, we computed the distance between the 3D eye-gaze vector in Vicon coordinates system and the hyper-plane containing the head of the interlocutor (picked up by reflector markers on the glasses). The distance values were then divided into two main groups, picture region or the interlocutor region. We defined a threshold separating the two groups for each trial. If distance sample was below the threshold, the sample was labelled as picture region otherwise, as interlocutor. From this analysis, we obtained the *fixation time to picture* and the *fixation time to interlocutor*.

Statistics

For each of the measures obtained from saccade and fixation analysis (i.e., number of saccades, fixation time to interlocutor and fixation time to picture), two linear mixed-effect (LME) models were executed. The first model described the effect of hearing status and included two fixed factors: hearing status (NH-N0 and SHL-N0) and repetition. The other model explained the effect of background noise with two fixed factors: noise level (NH-N0, NH-N60 and NH-N70) and repetition. The results were considered significantly different if $p < 0.05$. We ran the models with data samples extracted within the entire communication time (speaking time and listening time) as well as separately for speaking time and listening time.

Results

Saccades

Figure 1 illustrates that across the entire communication time, the number of eye-gaze saccades increased when conversing in noise (left panel) or with a SHL (right panel). The LME models confirmed the significant effect of noise ($F_{(2,120)}= 22.9$, $p< 0.001$) and hearing status ($F_{(1,72)}= 6.4$, $p= 0.01$) on the number of saccades. The significant difference between NH-N0 and two noise levels of NH-N60 ($p= 0.04$) and NH-N70 ($p<0.001$) was confirmed with a post-hoc pairwise comparison test.

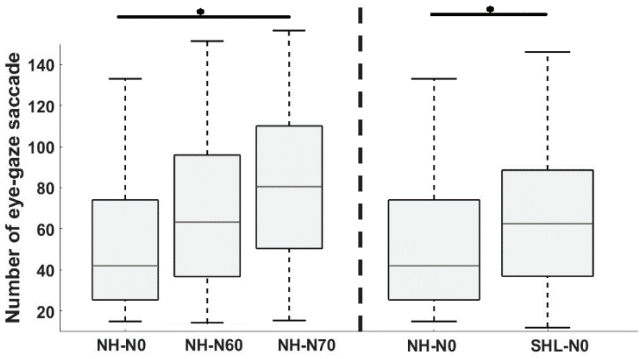

Figure 1: Number of eye-gaze saccades in two adverse conditions caused by noise on the left and by hearing status on the right.

Figure 2 shows the number of saccades measured in listening time and speaking time at different levels of background noise (left panel) and hearing status (right panel). LME models revealed a significant effect of noise in both speaking time ($F(2,110)= 17.4$, $p< 0.001$) and listening time ($F(2,110)= 6.9$, $p= 0.001$). The effect of SHL was significant only in the speaking time ($F(1,66)= 6.4$, $p= 0.01$).

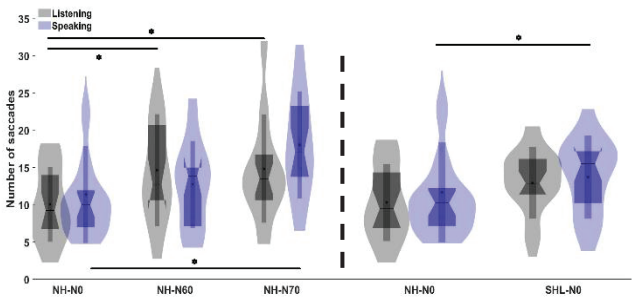

Figure 2: Violin plots of number of eye-gaze saccades across speaking time (blue) and listening time (gray) in two adverse conditions caused by noise (left panel) and hearing status (right

panel). The central asterisks show mean values, the central horizontal line is for median values, the rectangular shaded areas illustrate standard deviation, the dark shaded areas represent data boxplot and finally the large light shaded areas show kernel density plots.

Fixations

Figure 3 shows boxplots of the average fixation time to picture and interlocutor across communication time (combined speaking and listening time). Noise (right panels) and SHL (left panels) both caused a shorter fixation time on both the picture and interlocutor. The noise model as well as the hearing status model reported significant effects of background noise and hearing status on the average fixation time on both the picture (noise: $F_{(2,100)}= 4$, $p= 0.02$; hearing: $F_{(1,60)}= 11.3$, $p= 0.001$) and the interlocutor (noise: $F_{(2,100)}= 3.3$, $p= 0.04$ hearing: $F_{(1,80)}= 4.2$, $p= 0.04$).

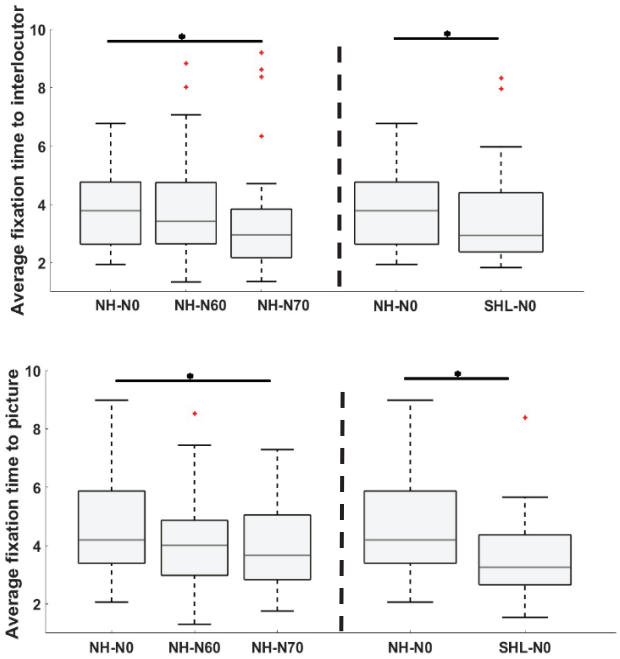

Figure 3: Average fixation time to the interlocutor (upper plot) and the picture (lower plot) in two adverse conditions caused by noise (left panel) and hearing status (right panel).

Figure 4 illustrates changes in the normalized fixation time on both ROIs within speaking time and listening time because of background noise and hearing status, respectively.

When fixating at the interlocutor in the listening time, there was a significant effect of noise level ($F_{(2,110)}= 5$, $p= 0.008$) but not of hearing status ($F_{(1,110)}= 2.2$, $p> 0.05$). When fixating at the interlocutor in the speaking time, there was a significant effect of both noise level ($F_{(2,110)}= 4.8$, $p= 0.01$) and hearing status ($F_{(1,66)}= 10$, $p= 0.002$).

When fixating at the picture in listening time, there was no significant effects of noise level or hearing status. The fixation at the picture decreased in the speaking time due to the effect of both noise level ($F_{(2,121)}= 3.6$, $p= 0.03$) and hearing status ($F_{(1,66)}= 5.1$, $p= 0.03$).

Discussion

In the current study, we investigated communication difficulty by looking into eye-gaze behavior during dyadic conversations. We investigated the total number of eye-gaze saccades and fixation times to two different ROIs (the picture and the interlocutor) at different background noise levels (NH-N0, NH-N60 and NH-N70) and at different hearing status (NH-N0 and SHL-N0). We considered those measures within the entire communication time but also separately within speaking and listening time.

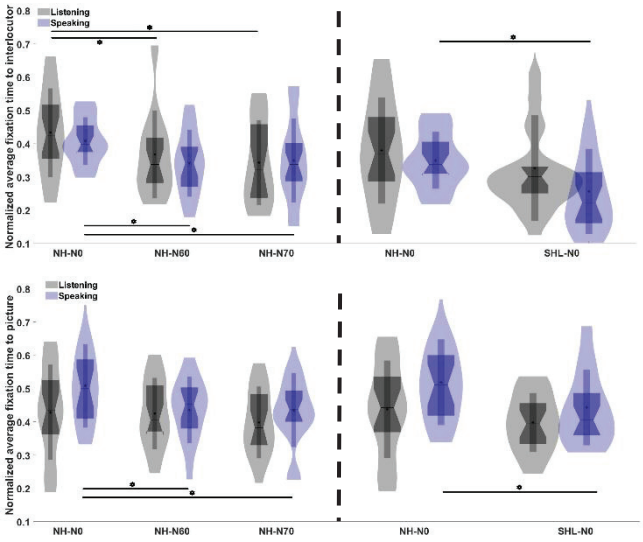

Figure 4: Normalized average fixation time to the interlocutor (upper row) and the picture (lower row) within speaking and listening time for different levels of background noise (left panels) and hearing status (right panels).

From the results on the entire communication time, we found that as communication gets harder, either by increasing noise levels or by introducing a mild conductive hearing loss, there is a significant increase in the number of saccades and a consequent significant decrease in the fixation time on both ROIs. This supports our hypothesis on more eye-gaze saccades as communication gets harder. The observed eye-gaze behavior is dependent on the task the participants are doing. To successfully complete the Diapix task participants need to access two sources of information: the picture and the interlocutor. As hearing get harder, then participants get more benefits from visual cues and hence this is reflected in more frequent changes in eye-gaze [13], [14]. Based on this, we believe that it is fair to conclude that the number of eye-gaze saccades and consequent fixation times can be used as potential indicators of communication difficulty caused by increased noise levels and increased hearing attenuation, when solving a Diapix task. We further suggest that these findings would extend to similar tasks like solving a logic puzzle [15].

To the best of our knowledge, there is no other study that has studied eye-gaze patterns while solving a similar task as Diapix. However, other studies have also identified systematic eye-gaze changes as communication difficulty increases. Hadley et al. did a similar experiment where they monitored eye-gaze when dyads hold a free talk conversation under different background noise levels [10]. While the authors could not see any changes on the total amount of fixations at different noise levels, they could see longer fixation times at the mouth region when noise level increased.

These findings support the idea of using eye-gaze as a potential indicator of communication difficulty during free topic conversations.

An interesting finding in our study is the differences observed on the fixation time to different ROIs when looking into speaking time and listening time. There were no significant differences on the fixation time to the picture when participants were listening under different noise levels and hearing status. A potential explanation could be assigned to a tactic that participants used when solving the Diapix task; when listening to the interlocutor talking about details in the Diapix picture, participants still need to look at their own picture to identify any potential difference. We also observed that the differences on the fixation time to each ROI within the entire communication time were mainly driven by the speaking time and not by the listening time. This can be explained by the signaling strategy that participants used to smooth the communication difficulties for their partner. We also see this accommodation pattern on speech production measures, i.e., NH participants speak louder and slower in communication with an unaided HI compared to an aided HI person [12].

Conclusion

In this work, eye-gaze behavior showed sensitivity to communication difficulty caused by changes in noise and hearing levels. It can confirm eye-gaze may be used as a proxy for communication difficulty.

References

[1] J. Riddiford, P. Enticott, A. Lavale, and C. Gurvich, "Gaze and social functioning associations in autism spectrum disorder: A systematic review and meta-analysis," *Autism Research*, vol. 15, May 2022, doi: 10.1002/aur.2729.

[2] J. Leppanen, F. Sedgewick, J. Treasure, and K. Tchanturia, "Differences in the Theory of Mind profiles of patients with anorexia nervosa and individuals on the autism spectrum: A meta-analytic review," *Neuroscience and Biobehavioral Reviews*, vol. 90. 2018. doi: 10.1016/j.neubiorev.2018.04.009.

[3] V. Best, E. Roverud, T. Streeter, C. R. Mason, and G. Kidd, "The Benefit of a Visually Guided Beamformer in a Dynamic Speech Task," *Trends Hear*, vol. 21, 2017, doi: 10.1177/2331216517722304.

[4] M. A. Skoglund, M. Andersen, M. M. Shiell, G. Keidser, M. L. Rank, and S. Rotger-Griful, "Comparing In-ear EOG for Eye-Movement Estimation With Eye-Tracking: Accuracy, Calibration, and Speech Comprehension," *Front Neurosci*, vol. 16, 2022, doi: 10.3389/fnins.2022.873201.

[5] M. M. Shiell, J. Høy-Christensen, M. A. Skoglund, G. Keidser, J. Zaar, and S. Rotger-Griful, "Multilevel Modelling of Gaze from Hearing-impaired Listeners following a Realistic Conversation," *bioRxiv*, 2022, doi: 10.1101/2022.11.08.515622.

[6] T. Foulsham, J. T. Cheng, J. L. Tracy, J. Henrich, and A. Kingstone, "Gaze allocation in a dynamic situation: Effects of social status and speaking," *Cognition*, vol. 117, no. 3, pp. 319–331, 2010, doi: https://doi.org/10.1016/j.cognition.2010.09.003.

[7] M. M. E. Hendrikse, G. Llorach, V. Hohmann, and G. Grimm, "Movement and Gaze Behavior in Virtual Audiovisual Listening Environments Resembling Everyday Life," *Trends Hear*, vol. 23, 2019, doi: 10.1177/2331216519872362.

[8] E. Šabić, D. Henning, H. Myüz, A. Morrow, M. C. Hout, and J. A. MacDonald, "Examining the Role of Eye Movements During Conversational Listening in Noise," *Front Psychol*, vol. 11, 2020, doi: 10.3389/fpsyg.2020.00200.

[9] H. Lu, M. F. McKinney, T. Zhang, and A. J. Oxenham, "Investigating age, hearing loss, and background noise effects on speaker-targeted head and eye movements in three-way conversations," *J Acoust Soc Am*, vol. 149, no. 3, 2021, doi: 10.1121/10.0003707.

[10] L. V. Hadley, W. O. Brimijoin, and W. M. Whitmer, "Speech, movement, and gaze behaviours during dyadic conversation in noise," *Sci Rep*, vol. 9, no. 1, 2019, doi: 10.1038/s41598-019-46416-0.

[11] R. Baker and V. Hazan, "DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs," *Behav Res Methods*, vol. 43, no. 3, 2011, doi: 10.3758/s13428-011-0075-y.

[12] A. J. M. Sørensen, M. Fereczkowski, and E. N. MacDonald, "Effects of Noise and Second Language on Conversational Dynamics in Task Dialogue," *Trends Hear*, vol. 25, 2021, doi: 10.1177/23312165211024482.

[13] O. Sandgren, R. Andersson, J. van de Weijer, K. Hansson, and B. Sahlén, "Coordination of gaze and speech in communication between children with hearing impairment and normal-hearing peers," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 3, 2014, doi: 10.1044/2013_JSLHR-L-12-0333.

[14] L. 'Skelt, "See what I mean : hearing loss, gaze and repair in conversation," The Australian National University , 2006.

[15] T. Beechey, J. M. Buchholz, and G. Keidser, "Eliciting naturalistic conversations: A method for assessing communication ability, subjective experience, and the impacts of noise and hearing impairment," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 2, 2019, doi: 10.1044/2018_JSLHR-H-18-0107.