Audio Engineering Society

Conference Paper

Presented at the 2022 International Conference on
Audio for Virtual and Augmented Reality
2022 August 15–17, Redmond, WA, USA

The Role of Lombard Speech and Gaze Behavior in Multi-Talker Conversations

Mark Dourado[1,2], Henrik Gert Hassager[1], Jesper Udesen[1], and Stefania Serafin[2]

[1]*GN Audio Research, DK-2750, Ballerup, Denmark*
[2]*Multisensory Experience Lab, Aalborg University, DK-2450, Copenhagen, Denmark*

Correspondence should be addressed to Mark Dourado (mdourado@jabra.com)

ABSTRACT

Effective communication with multiple conversational partners in cocktail party conditions can be attributed to successful auditory scene analysis. Talkers unconsciously adjust to adverse settings by introducing both verbal and non-verbal strategies, such as the Lombard effect. The Lombard effect has traditionally been defined as an increase in vocal intensity as a response to noise, with the purpose of increasing self-monitoring for the talker and intelligibility for conversational partners. To assess how the Lombard effect is utilized in multimodal communication, speech and gaze data were collected from four multi-talker groups with pre-established relationships. Each group had casual conversations in both quiet settings and scenarios with external babble noise. Results show that fifteen out of sixteen talkers exhibited an average increase in loudness during interruptive speech in all conditions with- and without external babble noise when compared to unchallenged sections of speech. Comparing gaze behavior during periods of a talkers own speech to periods of silence showed that the majority of talkers had more active gaze when speaking.

1 Introduction

The ability to communicate with multiple talkers in a noisy environment is one of the most powerful skills of the human auditory system. The capacity to distinguish between relevant and non-relevant auditory information, referred to as auditory scene analysis [1], solves what is known as the "cocktail-party problem" [2][3].

A typical consequence of this problem is the Lombard effect, initially conceptualized as an increase in vocal intensity as a nonspecific response to noise. Studies in effects of masking have identified speech features that change with acoustic noise, such as the fundamental frequency (F0), vocal intensity, an increase in duration of vowels, a shift in energy frequency bands and formant center frequencies (predominantly F1 and F2), and spectral tilting [4][5]. This is defined as Lombard speech. A study by Weber et al. showed that while the concept of a "full" Lombard effect exists, it is difficult to determine a significant increase of all parameters for different talkers reading sentences out loud in anechoic settings [5].

However, recent studies have redefined the Lombard effect as to include the visible speech content, such as articulatory facial movement and lip reading, and further extending it to involve gesture kinematics as well, relabeling it as a multimodal phenomenon [6].

When several talkers are engaged in the same conversation, the main mode of verbal interaction involves the rapid exchange of turns at talking [7]. When this happens, a transfer of who has “the floor” occurs, during which the acoustic signals produced by the parties involved might partially overlap or be separated by a silent gap [7][8][9]. In terms of modality, having access to visual information from a talker’s face has shown to enhance intelligibility [10], as well as the ability to resolve perceptual ambiguity in a noisy environment. Specifically for conversational turn-taking, both talkers and listeners produce non-linguistic visual cues indicating the end of a turn [11][12].

Mehra et al. describe four principal issues required to computationally solve the cocktail-party problem, namely: speaker separation, noise suppression, signal enhancement and intent detection [13]. They propose using an augmented reality (AR) platform and "multimodal, Ego-centric sensing", in which a network of sensors, visual data and biofeedback would enhance natural signals. They present the notion of context-aware systems that not only assess the surroundings of the user, but also the users own behavioral state and engagement within a conversation, in order to adapt and adjust to different situations.

This paper investigates conversational dynamics in multi-talker settings, by assessing the role of Lombard speech as a salient conversational cue within this context. More specifically, a hypothesis is formed as to whether or not Lombard speech occurs during attempted or successful floor transfers and if these occurrences change with increased babble-noise levels. Furthermore, initial steps are taken towards understanding gaze behavior in multi-talker conversations, with the purpose of discerning patterns that could relay listener intent. The motivation behind the research presented in this paper, is to move towards informing context aware speaker separation- and signal enhancement systems, of the relevance of auditory objects within an augmented environment - real or virtual. To examine this, 160 minutes of multi-talker conversation across four different groups and three different levels of external noise were recorded. During these sessions, head orientation and gaze behavior was tracked, and speech was recorded for each individual talker.

2 Methods

2.1 Participants

A total of sixteen participants were recruited for the experiment. The total sample size was comprised of eight male and eight female participants, aged between 22 to 61 (M = 31.8, SD = 11.7). All participants reported having no hearing impairment or diagnosed hearing loss, nor any visual impairment or uncorrected vision. All participants were collected in groups of four using a purposive sampling method, by interviewing one participant from each group (totaling four groups) and by introducing the criteria for participation, after which three other members would be selected by the aforementioned participant. The criteria for participation were: 1) Participants must be native Danish speakers. 2) Participants in each group must have engaged in a group conversation with the other members of their respective group. 3) Participants must not have diagnosed hearing loss, cognitive disability or suffer from vision-impairing afflictions.

2.2 Questionnaire

A two-part questionnaire was administered to participants before and after the experiment. The first half of the administered questionnaire, which participants filled out prior to the experiment, related to a "closeness" construct adopted from the "Inclusion of the Other in the Self" (IOS) pictorial task [14] (see Fig.1). Each participant rated their perceived closeness to each other individual group member on a scale from one to seven, for a total of twelve observations per group.

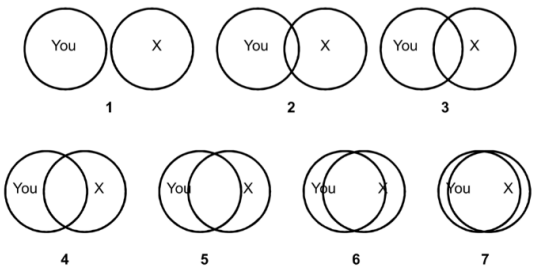

Fig. 1: IOS Pictorial Task.

The second half introduced two constructs, with the first including questions relating to the perceived naturalness of the conversations and the second relating to the dynamics of the conversations.

2.3 Setup

2.3.1 Hardware

Four Tobii Pro Glasses were used to track gaze behavior in addition to recording each wearers view through the head mounted camera, positioned at the bridge of the nose. Eight infrared Vicon Vero cameras were positioned around the inside of an elevated truss-frame. These were used to track head orientation and position via optical tracking of unique passive marker sets placed on each of the Tobii glasses. Four Genelec 8010A loudspeakers were placed equidistant around a round table, inside the tracked area. Four DPA 4066 Omnidirectional headset microphones were used to record the speech from each participant. All loudspeakers and microphones were routed into an RME Fireface 802 audio interface. All audio was played back and recorded at a sampling rate of 48 kHz.

2.3.2 Software

The Tobii Pro Glasses 3 Controller application, native to the glasses, was used to label and initiate recordings from the camera. Vicon Nexus data capture software was used to collect data from both the Vicon cameras and the Tobii glasses in the native environment. The Datastream SDK was used to simultaneously capture data from Vicon in Matlab. Matlab was used for all data analysis and post-processing (see 2.6).

2.4 Conditions and Stimuli

A within-subject test design was utilized, with three determined conditions, namely:

- Quiet - with no external stimuli
- 55 dBA babble noise
- 65 dBA babble noise

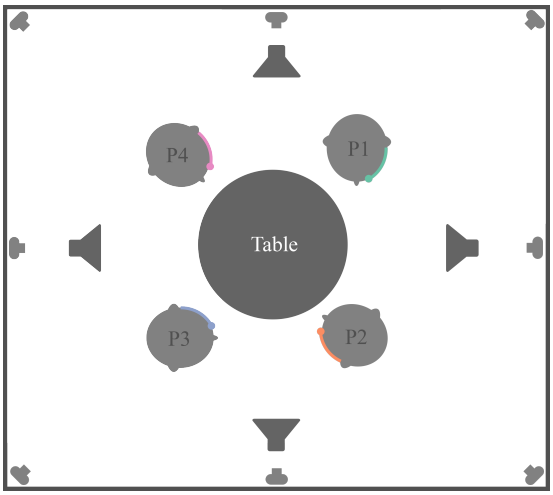

Fig. 2: Experimental setup. The eight shapes along the wall represent the position of each optical tracking camera.

The babble noise for the two respective conditions was presented over the four Genelec loudspeakers located 1.5 meters from the centre of the table, between each participant. The noise level was calibrated using a sound level meter at the centre of the table. Figure 2 shows the experimental setup with loudspeaker-, camera- , microphone-, and participant position around the table. The table had a diameter of 1 meter. The babble noise employed was created by using a frequency shaped babble noise generator with a pink-like spectrum [15]. Speech shaped noise from Dantale II [16] was initially tested at levels varying from 50 dBA to 80 dBA as the speech-competing noise for individual conditions; Similarly, Ambisonics recordings of cafe environments were tested at the same levels. Speech shaped babble noise was ultimately deemed the most ideal for the study as it would reduce the likelihood of participants being able to discern context or be distracted by transient occurrences in the soundscape. Furthermore, the noise would allow for "dip listening", i.e. the advantage gained from momentary improvements in the signal-to-noise ratio (SNR) of an acoustic scene [17]. Each of the four channels of noise were uncorrelated, as to avoid comb-filtering due to head-movement.

2.5 Procedure

The participants were asked to sit down prior to the beginning of the trial and were assigned a number, depending on their seat position. After choosing their

seat, the participants were escorted to different areas where they were asked to fill out the first half of a questionnaire. Each participant was re-seated and was fitted with a microphone and a pair of Tobii glasses, connected by an HDMI cable to a receiver placed underneath the table. The chairs were placed on markers denoting the placement of chairs during calibration of the system. Each participant was seated across another individual and approximately equidistant to their relative left- and right conversational partners. They were allowed to move around the chair, as to the extent allowed by its fixed position, i.e. leaning back, forth and sideways was possible and allowed. Before initiating the first recording, participants were made aware that the test conductor would appear multiple times to change between conditions and use the clapperboard to time-align the session. Participants were instructed to simply engage in casual conversation as they saw fit and were encouraged to continue any ongoing conversation between conditions. Figure 3 shows the seating arrangement and equipment.

Fig. 3: Seating arrangement and equipment. Clapperboard (middle of the table), four Tobii glasses, and DPA microphones (on the chairs).

All participants were subjected to each condition in the following sequence: Quiet (just four talkers with no external noise) - 55 dBA of external babble noise - 65 dBA of external babble noise - Quiet(2) (Quiet condition repeated), with each condition lasting 10 minutes, for a total of 40 minutes of recorded data. Every 10 minutes, a test-conductor would enter the room to change between conditions and start new recordings. For each initialization of a new condition and subsequent recordings, a clapperboard with tracking markers was used to create an impulse for alignment of audio, gaze and motion data. No formal breaks were introduced and participants were encouraged to keep the conversation going. However, sometimes a partial re-calibration was advised by the system and was thus performed between conditions. Participants remained seated throughout the recordings, and no equipment was rearranged or unequipped during the period. After the trial, the participants were instructed to fill out the second part of the questionnaire.

2.6 Post-processing

A normalized least mean squares (NLMS) algorithm was used to remove crosstalk from the four recorded microphone signals to obtain the individual speech signals for each of the four participants [18][19]. The noise signals played back over the loudspeakers were provided to the NLMS algorithm to remove the external babble noise from the recorded microphone signals.

The average of the top peak of the four impulses recorded from the clapper was labeled as time zero for the audio recording.

A voice-activation detection (VAD) algorithm was then used to label segments of active speech from the signal of each talker, resulting in a binary representation of speech onset and offset [20]. To validate the VAD, a signal containing only white Gaussian noise was multiplied with the VAD envelope and added to a copy of the respective speech signal, as to allow for visual inspection by plotting the spectrogram of the summed signal. This was done for all signals and incorrectly labeled segments were manually corrected.

Gaze vectors and head positions for each of the participants were streamed from Nexus to Matlab using the Datastream SDK provided by Vicon. The sample rate of the gaze vector and head position data was 200 Hz. At first, outliers in the gaze vectors were removed using a Hampel filter with a window of 16 adjacent samples on either side of the affected sample resulting in a sliding window of 33 samples. Samples deviating more than three standard deviations from the window median was replaced with the window's median. Secondly, missing samples were replaced by a linear interpolation using the two adjacent samples. Lastly, the gaze vector was smoothed using a one pole bi-directional recursive filter with a forgetting factor of 0.9. The outlier removal, interpolation, and smoothing was done in the euclidean space for each of the dimensions separately. The relative gaze angle to the other participants was

computed from the participant's processed gaze vector and the vector from their head position to the other participants.

The optical markers on the clapper were used to establish the point in time when the two arms would meet up and return to their closed positions, in order to determine time zero for the relative gaze angles.

A classification of who the participants were looking at was created by truncating the relative gaze angles to each of the other participants. This truncation threshold was set to 15 degrees, resulting in a total angle of 30 degrees for each other participant. The head covers approximately 12 degrees for a person with a head diameter of 18 cm at a distance of 150 cm.

The speech signals recorded from the microphones were calibrated using the blocks of speech for each individual (see section 2.7). The average level difference of the other microphones to the microphone attached to the individual of interest was found by using the speech signals recorded in the quiet condition, before the NLMS algorithm was applied.

2.7 Measures

The perceived loudness in phons was computed for each of the individual speech signals using the ISO 532-1 standard (Zwicker method) [21]. The perceived loudness was computed every 2 ms.

Based on the VADs, two categories of speech segments were created for each group, condition, and participant. Firstly, the blocks of speech where only one participant was talking were identified for each individual. The perceived loudness was then extracted for each of these blocks at every 50 ms and labeled as unchallenged speech (US) segments; This was done by taking the average of 25 consecutive samples of the perceived loudness. Next, the blocks of speech were located for each individual, where at least one other participant was talking for the same duration. Similarly, the perceived loudness was then extracted for these blocks and labeled as interruptive speech (IS) segments instead. To reduce incorrect labeling of backchannel responses as IS, only blocks larger than 600 ms were used.

The number of jumps in the participant's gaze were counted in the speech- and silence blocks (i.e. non-speech blocks). The duration of speech and silence was also found for each block.

Two measures were then computed, namely: the number of gaze jumps while talking per minute, and the number of gaze jumps while not talking per minute. Based on these two measures, a new measure was computed as the ratio of these. This was named the gaze jump ratio.

3 Results

3.1 Self-assessment

The closeness scores of all four groups were collected and are shown in Figure 4.

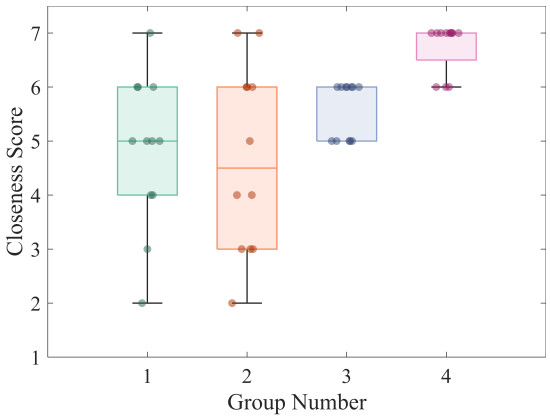

Fig. 4: Closeness scores of the individuals groups.

Using a Kruskal-Wallis one-way ANOVA, it was found that the perceived closeness between groups differed significantly ($\chi^2 = 9.52$, $p = 0.0232$). A post-hoc pairwise comparisons test showed that groups 4 and 2 were found to differ significantly ($p< 0.05$).

The items from the second half of the questionnaire are reported below. All responses (n = 16) are included for these. The scale ranges from 1 (strongly disagree) to 5 (strongly agree).

"I often have conversations with the group as a whole" (M = 4.56, SD = 0.73). "I was very aware of my surroundings during the conversations" (M = 2.81, SD = 0.98). "I was very aware of the equipment I was wearing during the conversations" (M = 3.31, SD = 1.14). "I was very aware that the conversations were being recorded" (M = 2.18, SD = 0.83). "The conversations felt unnatural" (M = 1.19, SD = 0.40). "I had to exert myself to take part in the conversation" (M = 1.25, SD = 0.45). "I was careful not to interrupt the others when they were talking" (M = 2.87, SD = 1.09). "The conversations were superficial" (M = 1.25, SD = 0.58).

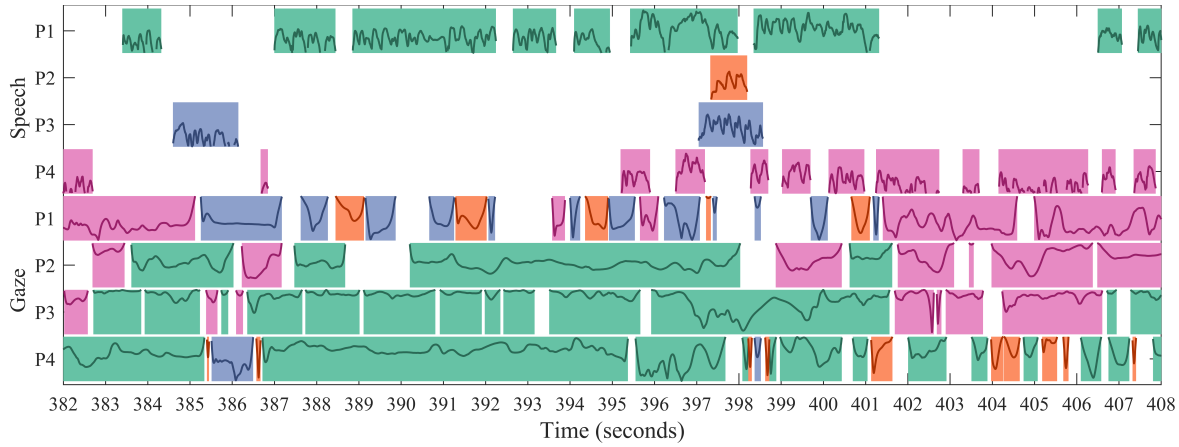

Fig. 5: An example of a conversation described by the perceived loudness and participant gaze: Group 3 in the 55 dBA babble condition from 382 seconds to 408 seconds. The top from rows show the four participants speech segments indicated in green, orange, blue, and pink. The perceived loudness is displayed on top of the speech segments. The perceived loudness ranges from 65 phons to 100 phons. The bottom four rows show the participants' gaze segments indicated in the corresponding colors of their fellow group participants. The relative gaze angle is displayed on top of the gaze segments. The gaze angle ranges from 0 degrees to 15 degrees.

3.2 Conversational Dynamics

Figure 5 shows speech activity and gaze behaviour and highlights different behavior and phenomena, such as:

- 382 seconds - 387 seconds: Conversation without crosstalk and a floor transfer from P1 to P3.
- 387 seconds - 395 seconds: P1 (green) has the floor. P1 has many gaze jumps while the others have few and are looking at the main talker throughout the duration.
- 395 seconds - 401 seconds: A floor transfer is attempted by P4 (pink). P1 increases their vocal intensity and so does P4, resulting in crosstalk with increased loudness.
- 401 seconds - 406 seconds: P1 yields their turn and transfers the floor to P4. The vocal intensity of P4 subsequently decreases along with an increase in gaze jumps during this duration.

3.3 Lombard Speech

A linear mixed model was used to determine the statistical significance of noise on loudness, for each participant. Results showed a significant main effect of noise for fifteen out of sixteen participants (p <0.05), for both 55dBA and 65dBA babble noise, consistent with the traditional Lombard effect. A participant with varying significance across conditions was participant 1 in group 3, between Quiet and Quiet(2) [F(3,1911) = -4.819, p<0.001], 55dBA babble and Quiet(2) [F(3, 1911) = .734, p=.256], 65dBA babble and Quiet(2) [F(3, 1911) = 2.152, p<0.001].

The perceived loudness between US segments and IS segments, for each participant and each condition were compared. Results showed that the majority of comparisons (64 total) featured higher average values of loudness for speech during IS segments than for US segments (98.44% for mean, 93.75% for median). 84.38% of US-IS comparisons have significantly different medians, with a 95% confidence interval (CI). For these comparisons, the difference in phons between the US CI upper bound (UB) and IS CI lower bound (LB) were calculated for each condition and participant. The second condition (55 dBA babble) had the largest average difference between US (UB) and IS (LB) (M = 5.07, SD = 2.66), with Quiet(2)(M = 4.29, SD = 2.80), Quiet (M = 3.63, SD = 2.17), and 65 dBA babble (M = 2.24, SD = 1.44), in order of largest to smallest.

A Wilcoxon Signed Rank test comparing US and IS

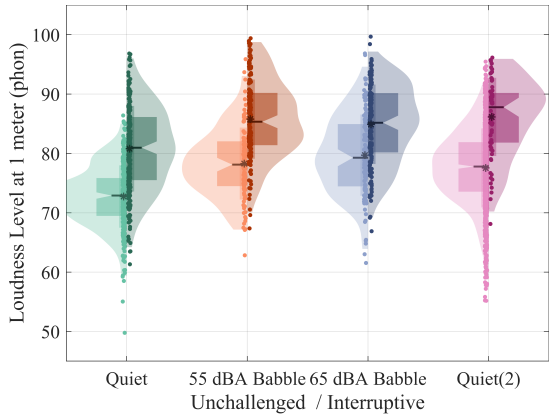

Fig. 6: Violin plots showing the perceived loudness for participant 1 in group 3. The perceived loudness is shown for the four conditions in green, orange, blue, and pink. The unchallenged speech segments are shown on the left and the interruptive speech segments are shown on the right. The star (*) inside the boxes denote the means and the black lines represent the medians. Notches depict the confidence interval of the median.

segments revealed that loudness scores were significantly higher for 89% of comparisons (p<0.05). An example from one of the comparisons where IS is significantly higher than US can be seen for participant 4 in group 4: Quiet (Z=-8.938, p<0.001), for 55dBA babble (Z=-8.302, p<0.001), 65dBA babble (Z=-7.740, p<0.001) and Quiet(2) (Z=-4.747, p<0.001).

Figure 6 shows the distribution of US and IS segments, and their perceived loudness in phons for one participant and all conditions. Figure 7 shows the distribution of US and IS for all participants.

3.4 Gaze Behavior

Figure 8 shows gaze jumps while talking per minute as function of gaze jumps while not talking per minute for all groups, conditions, and participants. It can be seen that all participants, except participant 3 in all groups, and participant 4 in group 1, have more gaze jumps while talking compared to when they are not talking. Figure 9 shows the gaze jump ratio as function of talking time per condition. Similarly, it can be seen that all participants, except participant 3 in all groups, and participant 4 in group 1 have a gaze jump ratio larger than one, meaning they have more gaze jumps while talking compared to when they are not talking. Furthermore, it can be seen that some participants (i.e. participant 4 in group 4) was talking for more than 50 % of the time (more than 5 minutes of the 10 minutes) and others less than 10 % of the time (i.e. participant 1 in group 3). It should be noted that Participant 1 in Group 4 in Condition Quiet(2) had a Gaze Jump Ratio of 2.53 for a talking time of 0.85 minutes.

4 Discussion

The closeness scores between each group very clearly demonstrated a bias of perceived closeness for some groups. This is likely due to the method of acquiring participants and the constellation of specific groups. Groups 1 and 2 consisted of more than one nominal category, e.g. group 1 was made up of four friends with two of them being married. Group 2 included four friends with two participants being twins. Group 3 included four friends who knew each other from college and group 4 was comprised of four family members. A future consideration would be to require that all participants have the same labels within each group.

The main reason behind the purposive sampling was to assess some of the potential advantages (and disadvantages) of more naturalistic conversations and group dynamics that could occur in a more ecologically valid context. One assumption was that in order to cultivate the observed increase in vocal intensity during crosstalk and conversational turn taking, pre-established and familiar group dynamics would make participants more comfortable with interrupting other conversational partners and streams of speech. Another assumption was that it could reduce the effect that the laboratory setting and equipment could have on the interaction within the groups. Nonetheless, the trade off between the exclusion of a conversational task meant that there was nothing to ensure equal floor time for every speaker. Had this been the case, the data sets would likely have been more balanced, but at the potential cost of reducing external validity or diminishing salient cues that would be relevant for study. The results are similarly limited by the ethnographic approach, as it would be logical to assume that conversational behavior (both gaze and speech-related) differ from culture to culture, and between generations of people. Thus, data from different populations would be required to produce more generalizable results.

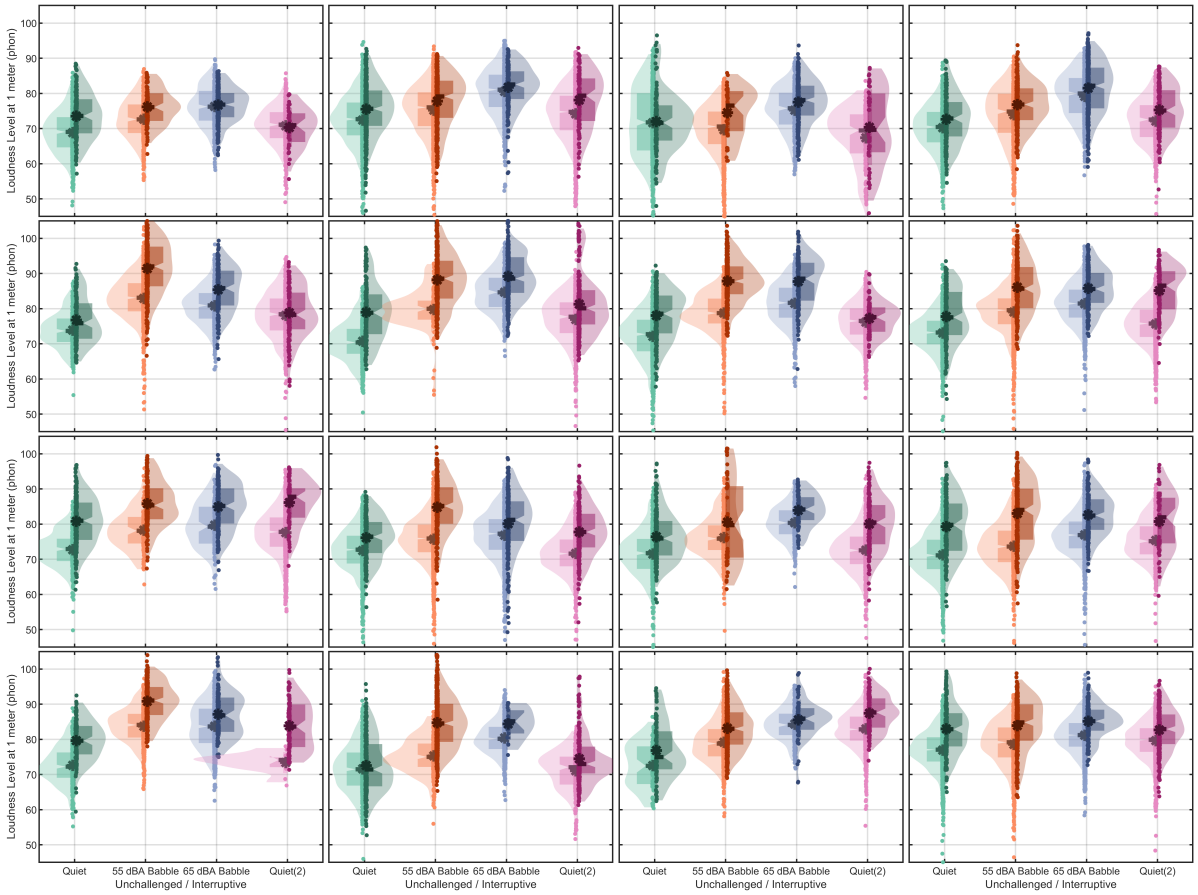

Fig. 7: The perceived loudness for all groups and participants. Groups in rows and participants in columns starting in the top left corner. The same order of colors is used denote the different conditions, as seen in Fig. 6.

However, for the groups that were represented in this study, results showed a general trend towards increased loudness for US segments with increased noise levels, consistent with the classical Lombard effect. Comparing US and IS segments, the majority of talkers exhibited a general increase in loudness in all conditions. Due to the test design, the unbalanced datasets meant that a few comparisons should be omitted when assessing trends, as the lack of speech in one condition would result in too little data (e.g. figure 7, group 4, participant 1). This is the results of an active session with very little uninterrupted speech or a passive session with little speech interaction whatsoever. It is also highly likely that the interaction is speaker dependant or that longer sessions of conversation would change the speech and gaze behavior. Lastly, manual labeling, more complex models and more data are required to robustly emphasize and validate the occurrence and categorization of conversational Lombard. While this study has predominantly focused on intensity as a way to represent Lombard speech, F0 estimation, harmonic-to-noise ratio (HNR), and SNR could be included for a more dynamic categorization. Lower level labeling of the conversational dynamics, such as categorizing floor transfer offsets and overlaps, backchannel utterences, and interpausal units (stretches of speech with silence less than 180ms) would only advance the ability to assess turn-taking desire or listener intent.

The same can be said for the categorization of gaze behavior. Most participants were shown to have had more active gaze when talking than when listening - a clear example of this was shown for one group in figure

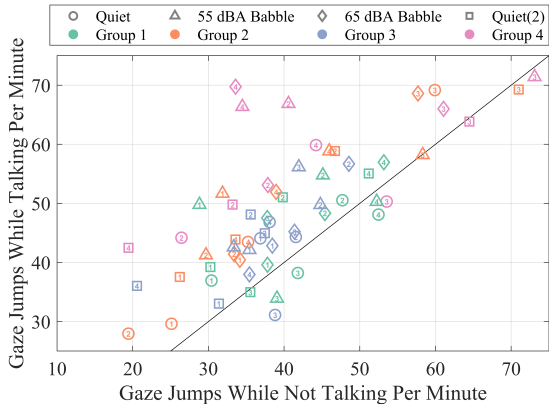

Fig. 8: Gaze jumps while talking per minute as function of gaze jumps while not talking per minute. The symbols are numbered to indicate each of the participant in the groups. The solid black line shows an equal amount of gaze jumps while talking and not talking.

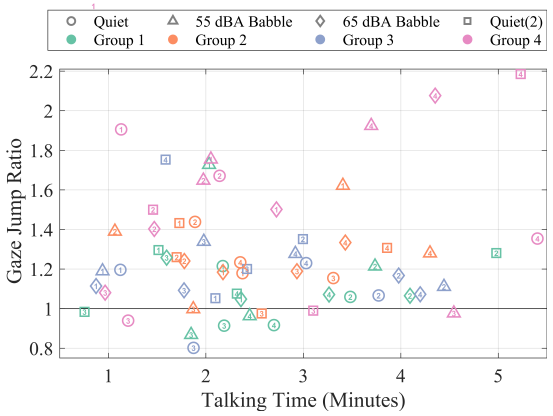

Fig. 9: Gaze jump ratio as function of talking time per condition. The symbols are numbered to indicate each of the participant in the groups. One outlier is omitted from the figure, namely participant 1 in Group 4 in Condition Quiet(2) had a Gaze Jump Ratio of 2.53 for a talking time of 0.85 minutes.

5. More categories for gaze as a function of speech (in this case, talking or not talking) would likely lead to more conclusive results. With that said, it also comes down to the design of the study, as dividing less than a minute of speech (see figure 9) into several smaller subdivisions leaves a very small sample size where any inferences made will be difficult to support.

5 Summary

In this study, four groups of people with pre-established relationships engaging in casual conversation were studied. A within-subject design was used and each group experienced a total of forty minutes of conversation with ten minutes allocated for each condition, namely: Two quiet conditions with no external noise and two noisy conditions featuring babble noise at 55 dBA and 65 dBA, respectively. Gaze-behavior, head orientation, position, and audio was recorded and collected for each participant. Audio was processed to reduce crosstalk from other talkers and a VAD algorithm was used to label segments of active speech for each individual talker. The relative gaze was thresholded and time aligned with the processed audio, resulting in segmented representations of speech and gaze for each participant. Comparing US- and IS segments, results suggest a trend towards increased loudness during interruptive speech in both conditions with- and without external babble noise. Comparing gaze behavior for periods of own speech and periods of silence for each participant, a proclivity towards more active gaze behavior for the person speaking could be seen for most participants, but not all.

Acknowledgements

We thank A. Josefine Munch Sørensen for providing us with great insight, fruitful conversation and much needed perspective. We also thank Clément Laroche for his technical assistance, his interest and his general eagerness to share and collaborate.

References

[1] Bregman, A., *Auditory scene analysis: The perceptual organization of sound.*, MIT Press, 1 edition, 1990.

[2] Cherry, E. C., "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, 25(5), pp. 975–979, 1953.

[3] Bronkhorst, A. W., "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Attention, Perception, and*

Psychophysics, 77, pp. 1465–1487, 2015, ISSN 1943393X, doi:10.3758/s13414-015-0882-9.

[4] Stowe, L. M. and Golob, E. J., "Evidence that the Lombard effect is frequency-specific in humans," *The Journal of the Acoustical Society of America*, 134, pp. 640–647, 2013, ISSN 0001-4966, doi:10.1121/1.4807645.

[5] Weber, D., Zaporowski, S., and Korzekwa, D., "Constructing a Dataset of Speech Recordings with Lombard Effect; Constructing a Dataset of Speech Recordings with Lombard Effect," 2020.

[6] Trujillo, J., Özyürek, A., Holler, J., and Drijvers, L., "Speakers exhibit a multimodal Lombard effect in noise," *Scientific Reports*, 11, 2021, ISSN 20452322, doi:10.1038/s41598-021-95791-0.

[7] Levinson, S. C. and Torreira, F., "Timing in turn-taking and its implications for processing models of language," *Frontiers in psychology*, 6, p. 731, 2015.

[8] Sørensen, A. J. M., MacDonald, E. N., and Lunner, T., "Timing of turn taking between normal-hearing and hearing-impaired interlocutors," in *Proceedings of the International Symposium on Auditory and Audiological Research*, volume 7, pp. 37–44, 2019.

[9] Sørensen, A. J. M., Fereczkowski, M., and MacDonald, E. N., "Effects of noise and L2 on the timing of turn taking in conversation," in *Proceedings of the International Symposium on Auditory and Audiological Research*, volume 7, pp. 85–92, 2019.

[10] Latif, N., Alsius, A., and Munhall, K., "Seeing the way: The role of vision in conversation turn exchange perception," *Multisensory Research*, 30(7-8), pp. 653–679, 2017.

[11] Jokinen, K., Furukawa, H., Nishida, M., and Yamamoto, S., "Gaze and turn-taking behavior in casual conversational interactions," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(2), pp. 1–30, 2013.

[12] Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., and Rubinstein, M., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, 37, 2018, ISSN 15577368, doi:10.1145/3197517.3201357.

[13] Mehra, R., Brimijoin, O., Robinson, P., and Lunner, T., "Potential of augmented reality platforms to improve individual hearing aids and to support more ecologically valid research," *Ear and hearing*, 41(Suppl 1), p. 140S, 2020.

[14] Gächter, S., Starmer, C., and Tufano, F., "Measuring the closeness of relationships: A comprehensive evaluation of the 'inclusion of the other in the self' scale," *PLoS ONE*, 10, 2015, ISSN 19326203, doi:10.1371/journal.pone.0129478.

[15] Pigeon, S., "Babble Noise Background Noise Generator," 2022.

[16] *Dantale II. Danske Hagerman Sætninger*, DTAS, 2001.

[17] Vestergaard, M. D., Fyson, N. R., and Patterson, R. D., "The mutual roles of temporal glimpsing and vocal characteristics in cocktail-party listening," *The Journal of the Acoustical Society of America*, 130(1), pp. 429–439, 2011.

[18] Chinaboina, R., Ramkiran, D., Khan, H., Usha, M., Madhav, B., Srinivas, K. P., and Ganesh, G., "Adaptive algorithms for acoustic echo cancellation in speech processing," *International Journal of Research and Reviews in Applied Sciences*, 7(1), pp. 38–42, 2011.

[19] Meyer, P., Elshamy, S., and Fingscheidt, T., "Multichannel speaker interference reduction using frequency domain adaptive filtering," *Eurasip Journal on Audio, Speech, and Music Processing*, 2020, 2020, ISSN 16874722, doi:10.1186/s13636-020-00180-6.

[20] Giannakopoulos, T., "A method for silence removal and segmentation of speech signals, implemented in Matlab," *University of Athens, Athens*, 2, 2009.

[21] ISO, "ISO 532-1: 2017. Acoustics—Methods for Calculating Loudness—Part 1: Zwicker Method," *International Organization for Standardization Geneva, Switzerland*, 2017.