INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Head movements in two- and four-person interactive conversational tasks in noisy and moderately reverberant conditions

Alan Archer-Boyd and Rainer Martin

Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany
{alanarcherboyd}@gmail.com, {rainer.martin}@rub.de

Abstract

Multi-modal processing schemes are of increasing importance for adaptive hearing devices. However, more data is required to understand interactions in complex application scenarios. In this study, the speech and head movements of eight normal-hearing participants were recorded in two- and four-person interactive conversational tasks, with and without 4-talker babble noise at 75 dB(A) and reverberation times of 0.25 and 0.6 s. Two-person conversations showed a head movement (yaw) interquartile range of 11.6° while four-person conversations showed a statistically significantly different interquartile range of 21.91°. No effect of acoustic condition was observed. The recorded data were also successfully used to test a previously published hearing-device direction of arrival estimation algorithm that utilized head movement information and correlation lag between acoustic signals from the left and right ear.
Index Terms: multi-modal processing, inertial measurement unit, source localisation, assistive hearing device

1. Introduction

Head movement is an integral part of human communication [1]. Therefore, real-time head movement information has the potential to greatly improve the performance of hearing-assistance devices. In order to properly investigate head movement during communication, it is important that experiments take place in fully interactive [2] and realistic situations [3].

In this study, we extend previous works towards conversations with up to four participants, and investigate the effect of background noise and different levels of reverberation in interactive conversational tasks. Head movement and bilateral behind-the-ear (BTE) hearing-aid microphone audio were recorded from normal-hearing participants in two- and four-talker interactive conversational tasks in quiet and babble and in low and moderately reverberant conditions. The interactive conversational tasks were chosen to represent two very different scenarios in terms of complexity. The two-person task was a game of "20 questions" which elicited mostly short questions and answers from the participants. The four-person task was a mission survival scenario, in which the group had to decide the relative importance of a list of items in a given emergency. By recording behaviour during these tasks, a direct comparison of the change in head movement and utterance behaviour could be made between simple, two-person and relatively complex four-person conversational tasks. Two noise levels were chosen (quiet and 75 dB(A)) to investigate the effect of adding a level of background noise comparable to loud restaurant noise, which is typically in the range of 60 – 80 dB(A) [4], [5].

[1]First author now with BBC Research & Development, UK.

2. Related work

The first attempts to quantify head movement during conversation were published in [6]. Head movement was measured quantitatively using a wall-mounted sensor and head-mounted markers. Talkers were found to move their heads more than listeners. Later work attributed certain types of movement to talker and listener intent, such as wide, linear movements ("postural shifts") that signified the initiation of speech and turn-taking. This suggested that head movement was involved in the regulation of turn-taking during conversations [7]. In the last decade, more emphasis has been placed on listener movement behaviour while actively listening. Brimijoin and colleagues recorded the movements of normal-hearing and hearing-impaired listeners while orienting towards sources in noise [8]. In similar later research, [9] measured head movement during localization of speech sources in background noise, and found a detrimental effect of background noise on localization accuracy and latency.

A number of studies have also investigated talker and listener behavior during conversations under realistic experimental conditions. [10] recorded speech, head movement, and eye movement during two-person semi-structured conversations in varying levels of background noise. They found that for every 1 dB increase in background noise level, speech level increased by 0.32 dB. Listeners also moved closer to one another, reduced the length of their utterances, and focused more on the talker's mouth. In a follow-up study using conversation triads as well as dyads in a similar experimental set-up [11], they found that listeners oriented to talkers more in babble than in speech-shaped noise. Listeners also optimized their head orientations more and took longer conversational turns in the triads than in the dyads. The type of task also has an effect, as spontaneous dialogue (e.g. "small talk") results in faster speech and longer utterances than conversations that are focused on solving a specific task [12], though other studies have found the opposite [13, 10]. The results of these studies suggest that the level of background noise can have an effect on speech utterances, conversation structure, and the head position of conversation participants.

In our study, we hypothesized that head position and movement would differ between the two- and four-talker conversations, and may depend on the level of reverberation. In addition, we tested the performance of a direction of arrival (DOA) algorithm using the head movement data.

3. Methods

3.1. Participants

Eight (four female) normal-hearing participants were recruited to the experiment. Six were native English speakers and two

10.21437/Interspeech.2023-1007

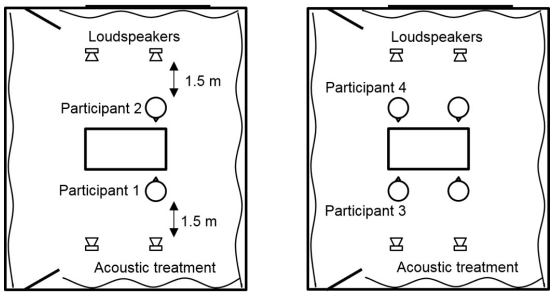

Figure 1: *Experimental set-up. Left: two-person set-up showing participant locations 1 and 2. Right: four-person set-up (participant locations 1 and 2 as annotated in the left plot)*

were fluent and non-native. Ethical approval was obtained for the collection of this data (ethics committee of RUB's medical department, file reference 5062-14). The participants had an average age of 29 years (range: 22-36). Their mean, four-frequency (0.5, 1, 2, 4kHz) audiometric thresholds were less than 25 dB HL in both ears. All participants took part in both the two- and four-person interactive conversational tasks. Two-person task pairs were gender balanced such that there were two male-female, one male-male, and one female-female pair. For the four-person tasks, participants were again individually assigned to two groups such that gender balance was maintained. Recordings took place over three sessions for each participant, one session for the two-person task and two sessions for the four-person task. The order of conditions was counterbalanced.

3.2. Experimental Set-up

Participants were seated in the middle of a lab room with adjustable acoustic curtains. Room dimensions were 7.57 m x 6.24 m x 2.90 m. Figure 1 shows a schematic of the set-up and the participant locations. Each participant was seated at the corner of a rectangular 1.6 x 0.8 m table, oriented along the short axis and facing another participant. The back of each chair was 0.4 m from the edge of the table, meaning that participants were approximately 1 m from their opposite conversation partner. The chairs were 0.45 m wide, meaning that the conversation partner on their side of the table was approximately 1.2 m from them. A list of words or survival items was placed in front of each participant, and participants familiarised themselves with the task before the recording session began. Participants were able to move freely while remaining seated.

3.3. Acoustic conditions

The participant conversation sessions were recorded in four acoustic conditions in a 2 x 2 design. Reverberation times (RT60) were measured using the average of eight recorded clapperboard "claps". These were (broadband average, 0.25-8kHz) 0.25 s (denoted "R1") with the acoustic curtains extended, or 0.6 s with the curtains drawn back ("R2"). Conversations took place in quiet ("N1"), or with four-talker babble presented at 75 dBA (measured from the participant's seating location) from four loudspeakers (one talker per loudspeaker) located 1.5 m behind the participants ("N2"). The babble consisted of four talkers (2 female) from the TSP Speech Database [14].

3.4. Tasks

3.4.1. Two-participant conversation

During the two-participant conversations, the participants were given a randomized list of household items, and took turns playing a game of "20 questions," in which one player could ask up to 20 questions in order to find out which item was next on the other participant's list. Roles reversed on each correct guess or when the answer was revealed. The number of questions asked was not important, and participants were informed beforehand that they could reveal the answer before it was correctly guessed if a particularly high number of questions had been asked. The game produced a minimal conversation turn-taking unit of a question and a yes or no answer and could easily be repeated. Each recording session lasted 10 minutes, and this was repeated for each acoustic condition and four pairs of participants, totaling 160 and 320 minutes of conversation and head movement recordings respectively.

3.4.2. Four-participant conversation

During the four-participant conversations, four variants of the NASA mission survival task [15] were used. Participants were given a list of 15 items and told that they were in a survival situation (e.g., stranded on the Moon, crash landed in a jungle). Their task was to discuss the items, and decide their order of importance for survival. In general, this provided sufficient conversational material for the length of the recording session. However, in the event that they finalized their list within the allotted time, they were encouraged to continue to converse more generally until the end of the session. The discussion of the items, in addition to most of the participants being strangers, also resulted in some digressions from the task. These were deemed acceptable and natural under the circumstances. Each recording session lasted 15 minutes, and this was repeated for each acoustic condition and two groups of participants, totaling 120 and 480 minutes of conversation and head movement recordings respectively.

3.5. Software and Hardware

Head movement was tracked using a 9-degrees-of-freedom (DOF) Razor inertial measurement unit (IMU) comprising an ITG-3200 gyroscope, ADXL345 accelerometer, and HMC5883L magnetometer. The device tracked individuals' yaw (angle in horizontal plane), pitch (median plane), and roll (frontal plane) movements without a fixed reference point.

Audio was recorded from eight pairs of BTE-mounted microphones (four per participant). Two participants wore Siemens micro-BTE hearing-aid chassis containing only their original microphones. One participant wore similarly constructed Siemens Acuris hearing-aid chassis. One participant wore custom devices comprising a front and rear-facing Sennheiser KE-4 microphone mounted on over-ear earphone guides in a similar position to the microphones on a BTE hearing aid. The equipment used varied as these were the only devices available to the researchers at the time of data collection. The use of different BTE devices would be unlikely to change participant behaviour or affect the conversation analysis. The audio sample-rate was 44.1 kHz. Four Logitech c270 HD webcams (Logitech Inc., Newark, CA, USA) were used to record each participant. Four talkers (two female) from the TSP speech corpus [14] were presented as background noise, each talker from one of four Genelec 2029BR (Genelec Oy, Iisalmi, Finland) loudspeakers placed 1.5 m behind each participant.

3.6. Annotation

Wavesurfer [16] was used to manually annotate the audio recordings from each recording session. These markers could be placed to an accuracy and minimum utterance length of 100 ms, and was based on the first author listening to the audio, and visually inspecting the waveform.

3.7. Analysis

Active talkers were marked manually. In line with previously published studies [17, 18], silent gaps of up to 1.25 sec between utterances by the same talker were marked as one continuous utterance. The 25th, 50th (median), and 75th quartiles of the position data (yaw, pitch, and roll) were calculated for each participant location in each condition, and each participant location while they were talking, or another participant was talking. For comparison across locations, the head position data was normalized and re-organized, such that a positive head position meant the participant was turning towards the participant at ±90°, and the individual participant listening data were classified as ahead (talker at 0°), diagonal (talker at ±40°) and adjacent (talker at ±90°). For example (see Fig. 1), in listener location 1, ahead was participant location 2, diagonal participant location 4, and adjacent participant location 3, whereas in listener location 3, ahead was participant location 4, diagonal participant location 2, and adjacent was participant location 1. Horizontal velocity was also analyzed between each time point in the same way.

A biomimetic DOA (BDOA) algorithm [19] was also applied to the data. 20 ms frames from the front left and right microphone signals from each participant were used as the inputs to a generalized cross-correlation with phase transform [20] algorithm. The peak in the cross correlation was converted to a direction of arrival relative to the head. The horizontal head position during each frame was subtracted and added from the DOA value to give the position of the source relative to the room for a source in the front or rear hemifield respectively, and these were plotted as histograms. Using this technique, robust absolute measures of multiple source positions can be built up over time without smearing the peaks in the histogram (and therefore the source position) due to head movement. Sources can be identified as being in the front or rear hemifield by picking the negatively or positively compensated histogram with the highest peak over a given time frame.

4. Results

4.1. Head positions during conversational tasks

Figure 2 shows the whole conversation mean median and mean inter-quartile ranges for yaw for each acoustic condition and conversation type (two- or four-person). Individual participant data (shown for the four-person conversations in Fig. 3) was highly variable. No clear pattern in the yaw position data could be seen across noise and reverberation conditions. It can be seen that there was little difference in either median or inter-quartile range in the two-person conversations. Median yaw across conditions was -1.05° and inter-quartile range was 5.87°. There was no significant difference between conditions in the four-person conversations, though there appears to be more variability in median and inter-quartile range. Median yaw across conditions was 11.06° and inter-quartile range was 21.91°. However, the mean median yaw and inter-quartile yaw ranges between the two- and four-person conversations were significantly different across participants (t(7) = -2.86, p=0.0243 and t(7) = -5.04, $p < 0.0015$, respectively, Bonferroni-Holm corrected).

Some variability was also observed in listener head pitch (not depicted), both in their median positions and their interquartile ranges. In the two-person conversations, median pitch was -2.38° (IQR: -5.54° to 0.31°) and roll was 1.45° (IQR: -2.85° to 6.68°). In the four-person conversations, median pitch position across all participants was -13.0° (IQR: -20.0° to -5.5°). Median roll position showed a similar pattern. Median roll position was -1.8° (IQR: -6.5° to 2.8°).

The head position data were analyzed during periods where the participants were speaking, or listening (defined as a period when another participant was talking). In the two-person conditions, median yaw position was similar whether talking (median = -1.5°; IQR: -5.8° to 2.9°) or listening (median = -1.0°; IQR: -5.0° to 3.0°). In the four-person conditions, the mean median position (yaw) across participants for periods speaking, and listening ahead, diagonal, and adjacent, were 13.5° (IQR: 3.4° to 24.7°), 8.0° (IQR: 0.5° to 17.3°), 15.5° (IQR: 3.6° to 24.7°), 15.9° (IQR: 3.8° to 38.5°) respectively. Two, two-tailed Student's t-tests (Bonferroni-Holm corrected) revealed a significant difference in head position between listening ahead and adjacent (t(14) = -2.82, p=0.014), but not for ahead and diagonal. The mean median positions when listening ahead in the two- and four-person conversations were significantly different across participants in both yaw (t(14) = -3.16, p=0.0069) and pitch (t(14) = 3.76, p=0.0021) (Bonferroni-Holm corrected).

4.2. Head-movement velocities during conversational tasks

The whole conversation distributions of head velocities in the two-person conditions were similar across participants and conditions. The distributions were symmetrical, and centered at or close to 0° s^{-1}. The proportion of time spent moving at a given absolute velocity decreased with increasing absolute velocity. Interquartile absolute velocity ranges were greater when participants spoke in comparison to when they were listening. This difference was significant in both yaw and pitch dimensions (yaw: t(14) = 3.25, p = .0058, pitch: t(14) = 4.09, p = .0017).

The whole conversation distributions of head velocities in the four-person conditions were similar across participants and conditions. The distributions were symmetrical, and centered at or close to 0° s^{-1}. The proportion of time spent moving at a given absolute velocity decreased with increasing absolute velocity. Interquartile absolute velocity ranges were greater for periods when participants spoke compared to listening, and there was no difference in velocity distribution between listening ahead, diagonal, or adjacent. These results were the same for both yaw and pitch. Bonferroni-Holm corrected Student's t-tests revealed that the effect of speaking vs listening was significant in both yaw and pitch (t(14) = 6.26, p<0.001, and t(14)= 5.65. p<0.001, respectively). The standard deviation of the interquartile ranges was also larger for speaking than for listening, showing that there is more variation in how fast participants move their heads while speaking than while listening.

4.3. Multi-modal DOA estimation

The application of a previously published multimodal BDOA algorithm to data collected from a real conversation serves as an example of the improvements to currently available algorithms by including head movement information. Conditions R1N1 and R2N1 were used (i.e. no background noise). Figure 4 shows that in the majority of recordings, the BDOA provided a better estimate of source position than using only the DOA, and in

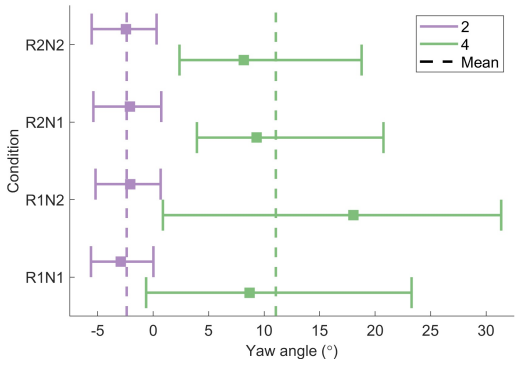

Figure 2: *Across-participant mean median (squares) and mean inter-quartile ranges (error bars) for each acoustic condition (y-axis) and number of conversation participants (colors). The x-axis shows the yaw angle. The dashed lines show the mean median across acoustic conditions.*

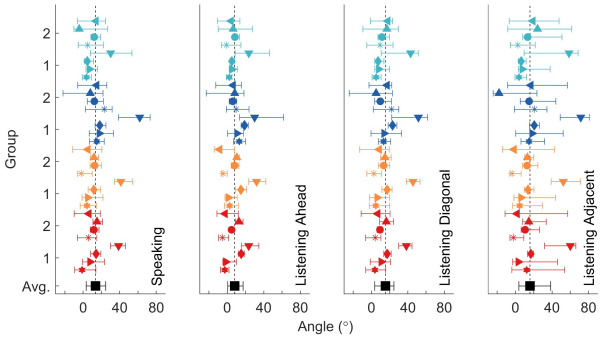

Figure 3: *Yaw head positions of individuals in each conversation and in four acoustic conditions (cyan: R1N1, dark blue: R1N2, orange: R2N1, red: R2N2). Data is shown for "speaking", "listening ahead", "listening diagonal", and "listening adjacent". 25th, 50th (median), and 75th yaw position percentiles (colors) for each participant (symbols). The black vertical dashed lines show the mean median values for each plot.*

all cases the BDOA correctly determined that the sources were in the front rather than rear hemifield (higher peaks in -BDOA than +BDOA). The application of the BDOA and the resulting source position predictions also served as a "sanity check" on the quality of the audio and head-position data, as poor alignment in time between the audio and head tracker samples, or poor calibration of the head trackers, would have resulted in the BDOA failing to predict the position of the audio sources.

5. Discussion

The two-person interactive conversational task was designed to provide a baseline, unscripted communication task that could be repeated across different acoustic conditions. The quartile analyses showed that there was no significant difference between the conditions and revealed the high variation between participants, in both two- and four-person conversational tasks. It also showed that head position was significantly different when participants were listening ahead in the two- and four-person tasks. This is evidence of a change in head position due to the conversational task. In the four-person task, participants turned more towards the middle of the table. There was also a significant

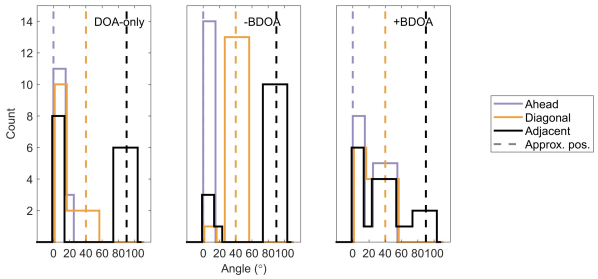

Figure 4: *A histogram of the peaks (absolute angle) in the individual DOA histograms after no correction of audio-derived DOA for head position (column 1, DOA-only), subtraction of head position from audio-derived DOA estimate (column 2, -BDOA), and addition of head position to audio-derived DOA (column 3, +BDOA). Estimates when the talker ahead was active are shown in teal, diagonal active in red, and adjacent active in black. The dashed lines in each color show the approximate position of the talker relative to the participant.*

difference in head position between listening ahead and listening adjacent, but not between ahead and diagonal, though the mean medians for diagonal and adjacent were similar (mean median yaws of 15.5° and 15.9° respectively). This suggests that there is one identifiable head position for on-axis listening, and another for off-axis listening, regardless of the angle of the off-axis source. One of the most robust findings across participants, conditions, and group sizes was the difference in head movement velocity distributions when participants were talking or listening. Consistently, in both yaw and pitch, the interquartile range of velocities was wider when participants spoke versus when they listened. This replicates a previous much earlier finding in quantitative conversation analysis [6], [7], and could provide an own voice detection cue for hearing-device algorithms that is robust to acoustic noise. The low number of participants was a clear limitation of this study. Eye movement data may have enhanced the findings, especially in the noisy conditions [10].

6. Conclusions

The speech and head movements of participants in two- and four-person conversations were recorded with and without background noise, and in rooms with two different reverberation times. Between-participant variation on almost all metrics was large. Participants moved their heads consistently more slowly when listening than when talking. Quartile analyses of head position showed large variations in head position across participants and conditions. A clear difference in head position was measured between the two- and four-person conversations when the talker was ahead of the participants: in the two-person conversations, participants oriented straight ahead, whereas in the four-person conversations, participants oriented towards the middle of the table. The angle of orientation only increased significantly when the talker was adjacent to the participant. No effect of reverberation or background noise was found on conversation head position. A source DOA algorithm that utilized head-position information was successfully applied to the data, showing the potential that head-movement information has for hearing-device algorithms.

7. References

[1] A. Kendon, "Some uses of the head shake," *Gesture*, vol. 2, no. 2, pp. 147–182, 2002.

[2] H. De Jaegher, E. Di Paolo, and S. Gallagher, "Can social interaction constitute social cognition?" *Trends in cognitive sciences*, vol. 14, no. 10, pp. 441–447, 2010.

[3] L. Verga and S. A. Kotz, "Putting language back into ecological communication contexts," *Language, cognition and neuroscience*, vol. 34, no. 4, pp. 536–544, 2019.

[4] C. P. Lebo, M. Smith, E. Mosher, S. Jelonek, D. Schwind, K. Decker, H. Krusemark, and P. Kurz, "Restaurant noise, hearing loss, and hearing aids," *Western Journal of Medicine*, vol. 161, no. 1, p. 45, 1994.

[5] L. H. Christie, "Psycho-to-building acoustics: are bars, cafes, and restaurants acceptable acoustic environments?" *Victoria University of Wellington*, 2004.

[6] U. Hadar, T. Steiner, E. Grant, and F. C. Rose, "Kinematics of head movements accompanying speech during conversation," *Human Movement Science*, vol. 2, no. 1-2, pp. 35–46, 1983.

[7] U. Hadar, T. J. Steiner, and F. C. Rose, "Head movement during listening turns in conversation," *Journal of Nonverbal Behavior*, vol. 9, no. 4, pp. 214–228, 1985.

[8] W. O. Brimijoin, D. McShefferty, and M. A. Akeroyd, "Auditory and visual orienting responses in listeners with and without hearing-impairment," *The Journal of the Acoustical Society of America*, vol. 127, no. 6, pp. 3678–3688, 2010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20550266https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4338612/

[9] A. Archer-Boyd, J. Holman, and W. Brimijoin, "The minimum monitoring signal-to-noise ratio for off-axis signals and its implications for directional hearing aids," *Hearing Research*, vol. 357, pp. 64–72, 2018.

[10] L. Hadley, W. Brimijoin, and W. Whitmer, "Speech, movement, and gaze behaviours during dyadic conversation in noise," *Scientific Reports*, vol. 9, no. 1, pp. 1–8, 2019.

[11] L. V. Hadley, W. M. Whitmer, W. O. Brimijoin, and G. Naylor, "Conversation in small groups: Speaking and listening strategies depend on the complexities of the environment and group," *Psychonomic Bulletin & Review*, pp. 1–9, 2020.

[12] S. Watson, A. J. M. Sørensen, and E. MacDonald, "The effect of conversational task on turn taking in dialogue," in *Proceedings of the International Symposium on Auditory and Audiological Research*, vol. 7, 2019, Conference Proceedings, pp. 61–68.

[13] T. Beechey, J. M. Buchholz, and G. Keidser, "Measuring communication difficulty through effortful speech production during conversation," *Speech Communication*, vol. 100, pp. 18–29, 2018.

[14] P. Kabal, "TSP speech database," Department of Electrical & Computer Engineering McGill University, Tech. Rep., 2002.

[15] J. Hall and W. H. Watson, "The effects of a normative intervention on group decision-making performance," *Human relations*, vol. 23, no. 4, pp. 299–317, 1970.

[16] J. Beskow and K. Sjolander, "Wavesurfer," https://sourceforge.net/projects/wavesurfer/.

[17] E. Campione and J. Véronis, "A large-scale multilingual study of silent pause duration," in *Speech prosody 2002, international conference*, 2002.

[18] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[19] A. Archer-Boyd, W. Whitmer, W. Brimijoin, and J. Soraghan, "Biomimetic direction of arrival estimation for resolving front-back confusions in hearing aids," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. EL360–EL366, 2015.

[20] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.