Who am I speaking at?
Perceiving the head orientation of speakers from acoustic cues alone

Jens Edlund[1], Mattias Heldner[2], Joakim Gustafson[1]
[1]KTH Speech, Music and Hearing, Stockholm, Sweden
[2]Linguistics, Stockholm University, Stockholm, Sweden
E-mail: edlund@speech.kth.se, mattias.heldner@ling.su.se, jocke@speech.kth.se

Abstract

The ability of people, and of machines, to determine the position of a sound source in a room is well studied. The related ability to determine the orientation of a directed sound source, on the other hand, is not, but the few studies there are show people to be surprisingly skilled at it. This has bearing for studies of face-to-face interaction and of embodied spoken dialogue systems, as sound source orientation of a speaker is connected to the head pose of the speaker, which is meaningful in a number of ways. We describe in passing some preliminary findings that led us onto this line of investigation, and in detail a study in which we extend an experiment design intended to measure perception of gaze direction to test instead for perception of sound source orientation. The results corroborate those of previous studies, and further show that people are very good at performing this skill outside of studio conditions as well.

1. Introduction

Gaze and head pose shifts are central to studies of human face-to-face interaction. They are becoming equally important for spoken dialogue systems research as the interest for embodied systems keep increasing. At the same time, the way in which we collect and use our corpora is changing. More and more corpora are not only multimodal in the traditional sense of containing both audio and video, but also hold other information such as movement data. And increasingly, we study dialogue within the situation: we attempt to model not only the dialogue itself and its semantic context, but facts about the space in which it takes place, about the moods and motivations of its participants, or about the events taking place in its vicinity. Finally, a growing community of researchers focus on developing spoken dialogue systems that are first and foremost humanlike, either because they are convinced that humanlikeness will improve spoken dialogue as a human-machine interface, or because they are interested in testing their hypotheses about how human interaction works.

In light of this altogether more holistic view of dialogue research, we have investigated the extent to which a listener can perceive a speaker's facing angle under normal conversational circumstances. To the extent that human speakers' facing angles are important, the auditory perception of a speaker's facing angle is likely to be important as well. We start out with a background of the area and of related research that serves as motivation for this study, continue with a brief description of the preliminary mini-studies that led us to the present study, and conclude with a detailed description of the present study, its method and its results.

2. Background and related work

The spatial relation between speakers and listeners is an important part of the dynamically changing situation in which a conversation unfolds. This spatial relation can be modelled using corpora in which reliable data describing each participant's orientation and location in the room is available from for example motion capture, such as the Spontal database (Edlund et al., 2010). Modelling the acoustic effects of these spatial relations would require, minimally, the addition of binaural microphones in each participant's ears. No available and sizeable corpus to date holds both binaurally captured sound and positional data.

2.1 Perception of sound source orientation

Whereas studies of people's ability to judge the position of a sound source are plentiful, there are only a handful studies of our ability to judge the orientation of directional sound sources.

In the early 2000s, Neuhoff and colleagues showed that people can indeed distinguish between different orientations of a directional loudspeaker. Neuhoff (2001) shows subjects' ability to detect the facing angle of a loudspeaker playing recorded speech in an empty room, and find that factors influencing this ability include whether the sound source is stationary or rotating (the movement helps); the distance to the sound source (closer is better); and the facing angle itself (the task is easier when the loudspeaker faces the listener straight on). Neuhoff et al. (2001) determines a just noticeable difference (JND) for facing angles by having subjects judge the orientation of a loudspeaker producing broadband noise in an anechoic chamber. As predicted by the findings in Neuhoff (2001), the JND varies with the distance to the loudspeaker and with the facing angle itself. The work is brought together and discussed in Neuhoff (2003), where greater weight is given to the bearing of these results on spoken interaction research. Neuhoff and colleagues implicate the inter-aural level difference (ILD) as the most likely cue to sound source orientation.

Kato and colleagues later took the potential relevance for realistic human-to-human telecommunication as their main motivation to perform similar studies. Kato et al. (2010a) and Kato et al. (2010b) both report on a study where a male speaker poised on a pivot chair in an anechoic chamber speak utterances at different horizontal

and vertical angles. We focus on the horizontal angles here. 12 blindfolded listeners were asked to indicate the speaker's facing direction. The results, including an average horizontal error of 23.5 degrees, are comparable to or better than those achieved with loudspeakers, adding evidence to the idea that interlocutors may be able to hear the head pose of the speaker from acoustic cues alone. A clear effect of the facing angle was observed, with head-on utterance being much easier to judge correctly. Kato and colleagues also analyse the acoustic transfer function from a speaker's mouth to the ears of a listener using binaural microphones, and like Neuhoff and colleagues, they find ILD to be the prime cue for horizontal orientation.

Finally, Nakano et al. (2008) and Nakano et al. (2010) contributed a comparison between perception in what they term a *real environment* - a normal room stripped bare of all furniture - and an anechoic chamber. Their stimuli is a live human speaker. Their subjects do better in the anechoic chamber. They also compare performance before and after a training session, and get an improvement from training.

2.2 Sound source orientation and face-to-face interaction

It is well attested that gaze, and in particular mutual gaze is important for the interaction in face-to-face dialogue. A typical gaze pattern, at least in Europe and in Northern America, is that the listener looks fairly constantly at the speaker, while the speaker looks at the listener in the vicinity of speaker changes or backchannels (e.g. Bavelas & Gerwing, 2011; Kendon, 1967). Hence, auditory perception of speaker facing direction might provide a redundant correlate of gaze in visible conditions, and a correlate of gaze in non-visible face-to-face conditions, such as in the dark. Note also, as mentioned above, that several studies report that listeners are particularly sensitive when the sound source is directed straight at them, that is, the situation correlated to mutual gaze in visible conditions.

2.3 Sound source orientation and embodied spoken dialogue systems

Currently, there are no interactive systems that detect and make use of sound source orientation, and systems that use gaze and head pose as a part of their expressive repertoire routinely produce audio through fixed loudspeakers without concern for what the acoustic effects of the head movements they display would be. Nakano et al. (2010), however, show a machine trained on acoustic data from an array microphone that perform better than chance but poorer than human subjects on the task of detection the facing angle of a speaker.

Given the importance of gaze in face-to-face interaction, there is considerable scope for improving the interactional capabilities of interactive avatars and robots by endowing them with means to produce and perceive visible as well as audible facing direction.

3. Preliminary studies

The idea that speaker head orientation may be heard by listeners struck us for no good reason during a conversation about turntaking a number of years ago. The thought immediately fascinated us, and we immediately proceeded to run impromptu tests and to track down and read up on the work of Neuhoff and colleagues, but time constraints came in the way of proper replication and publication. The tests we did run had a few things in common. They tested five orientations only - head on towards the listener, and 45 as well as 90 degrees in either direction. We felt that those directions were sufficient to study the effects the acoustics of face orientation might have on spoken face-to-face interaction. We used a real human speaker reading a predefined sentence, sacrificing the control afforded by a recording replayed in a directional loudspeaker for the ecological validity of a real human speech production apparatus. Tests in a number of environments, including offices, snow-clad fields and noisy bars, and at distances ranging from 1 metre to 10 metres all showed that subjects were able to indicate the direction in which speaker was facing from listening only with an accuracy was much above random choice. As we have recently increased our studies of co-presence (Edlund et al., 2011) as well as our efforts to create situated and embodied conversational partners (Al Moubayed et al., in press), we decided to resume these studies and repeat these tests under more controlled circumstances. And while the studies published to date were all performed in studios or rooms designed to minimize or normalize echoes, we choose to focus on a real everyday environment, sacrificing control for ecological validity.

4. Method

4.1 The subject/target experimental paradigm

We employed an experimental paradigm first used in Beskow & Al Moubayed (2010), where it was developed to allow experimenters to quickly gather large amounts of data on human perception of gaze targets/direction. We have generalized the paradigm here, and adapted it to work for perception of directional audio. In its generalized form, the paradigm is used to gauge subjects' ability to perceive the intended target of a directional stimulus, and can be described as follows.

A group of N subjects are placed in a circle or semi-circle, so that there is one point at their centre which is equidistant to each subject, from which all stimuli are presented (the *centre*). Subjects positions are numbered P_1 to P_N, and the angle between each subject's position, that of the centre, and that of the subject's closes neighbouring subjects ($A(P_1P_{2)}$... $A(P_NP_{1)}$) is calculated. Subjects may or may not be equidistant from their closest neighbours.

All subjects double as targets for the directional stimuli (hence the *subject/target paradigm*). During an experiment, directional stimuli are aimed at each of the subjects. The order is varied systematically, and the number of stimuli is such that each subject is targeted as

Figure 1: The experiment environment

many times as the others in one set of stimuli. A set of stimuli, then, contains a multiple R of N for a total of R*N stimuli. Once one set is completed, the subjects rotate - they shift their positions by one step and the process of presenting a set of N*R stimuli is repeated. The rotation is repeated N times, until each subject has been in each position once, making the total number of stimuli presented in an experiment N*R*N.

Each time a stimulus has been presented, each subject is asked to point out the intended target in such a manner that the other subjects cannot see it. The result is N judgements for each stimulus, for a total of N*R*N*N data points in one experiment. If more than one experiment condition is to be tested, the entire process is repeated from the beginning.

We now turn to the specifics of the present experiment.

4.2 Subjects

Two conditions were tested in a between-group design, and groups with five participants (N=5) were used. The subjects were students and university employees. Four of the subjects were female and six were male. All reported having normal hearing on both ears.

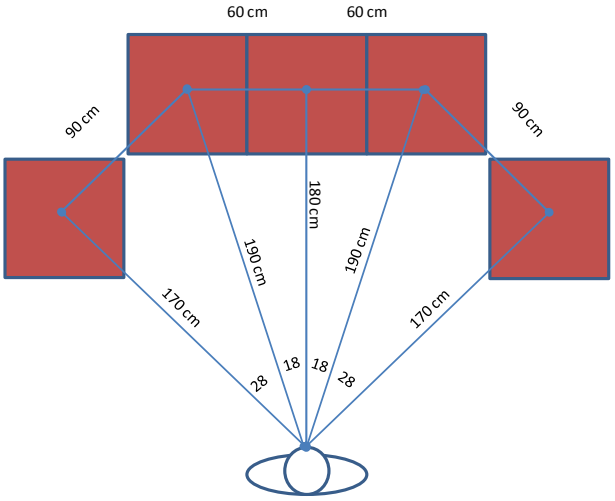

Figure 2: Schematic view of the experimental setup

4.3 Spatial layout and surroundings

The main motivation for the experiment was to test the subjects' ability to perceive acoustic (speech) directionality in normal, everyday conditions. For this reason, an existing recreational sofa group in busy office surroundings was chosen, and no attempt were made to stop other people from walking through the area or talking nearby. The sofa group was left standing as it is normally, and subjects were seated in five of the seats, as seen in Figure 1. A result of this was that the distance to the nominal "centre" from which stimuli were presented was not identical for all seats. The actual measures are shown in Figure 2, which also shows the distances and angles between adjacent subjects.

4.4 Stimuli

The experiment conductor spoke the sentence "Who am I speaking to now", while facing one of the subjects head-on from the nominal centre position. Each group contained two readings directed at each target (R=2) for a total of ten readings, after which the subjects were rotated.

4.5 Conditions

A between-group design was employed, in which the first group (NOFEEDBACK) were presented with stimuli exactly as described above, while the second group (FEEDBACK) received feedback after each utterance, once all five judgements had been recorded. Feedback consisted of the reader saying "I was talking to number N", where N was a number between 1 and 5 referring to the five seats from left to right. The subjects in this group had been informed about this procedure beforehand.

4.5 Responses

The subjects used hand signs to show which listener they thought the reader was facing: one, two, three or four fingers on the left hand to signify one, two, three and four steps to the left, respectively; one, two, three or four fingers on the right hand to signify one, two, three and four steps to the right; and a pointing gesture towards the chest to signify themselves (see figure 3).

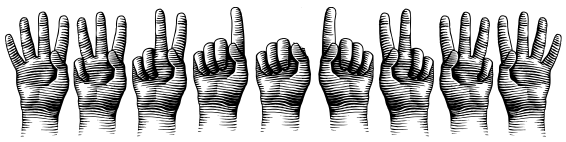

Figure 3: Signs used to indicate target position

All in all, the utterance was spoken 5*2*5=50 times for each condition. With five responses for each utterance, a total of 500 judgements were collected, 250 for each group and condition.

5. Results

Combined over the two conditions, the subject got the target exactly right in 259 out of 500 cases, or 52 % of the time. Random choice yields a 20 % baseline, and chi-square test shows that the result deviates significantly from a random choice ($\chi^2(1, N=24)=488.79, p=0.0001$).

The confusion matrix for all data is shown in Table 1.

Variance analysis of the errors (ANOVA), assuming equidistant positions, show significant main effects for condition, with the FEEDBACK condition resulting in a smaller error ($F(1,496)=4.23; p=.04$). No main effects

were found for gender *(F(1,496)=0.23; p=.63)*, nor were there any interactions between gender and condition.

Table 1. Confusion matrix for all subjects and conditions.

Target position \ Estimated target position	1	2	3	4	5	Total
1	**62**	23	7	8	0	**100**
2	13	**40**	38	8	1	**100**
3	9	15	**47**	24	5	**100**
4	1	8	37	**35**	19	**100**
5	0	2	6	17	**75**	**100**
Total	**85**	**88**	**135**	**92**	**100**	**500**

6. Discussion and future work

The results of the present study show that listeners are quite good at distinguishing between different facing angles in a speaker not only in anechoic chambers and emptied out, silent rooms, but also under conditions in which conversations normally occur - in furnished, asymmetric rooms with background noise and people passing by. This is consistent with an idea that the acoustic properties of speech and facing angle may be a redundant cue that interlocutors take into consideration in face-to-face spoken interaction. We further argue that modelling the acoustic properties of speakers' position and orientation is an important step in achieving a realistic model of situated interaction.

The data (see Table 1) also indicate that some directions in our fully furnished environment were easier to detect than others. This suggests that listeners use more than ILD to judge the facing angle of a speaker, but rather maintain an model of their acoustic environment into which they fit acoustic stimuli. As an example, when the speaker faced straight towards the large window set on his right side, subjects on all seats were more likely to judge the direction correctly, possibly due to the special acoustic character of the reflection against the window. This leads us to our next goal: to compare listeners' performance in everyday environments to anechoic chambers. If models of the acoustic environment are involved, one might expect poorer performance in an anechoic chamber; if it is all IDL, the anechoic chamber should instead help.

7. Acknowledgements

This work was funded by the Riksbankens Jubileumsfond (RJ) project *Prosody in conversation* (P09-0064:1-E).

8. References

Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (in press). Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. To be published in Esposito, A., Esposito, A., Vinciarelli, A., Hoffmann, R., & C. Müller, V. (Eds.), *Cognitive Behavioural Systems. Lecture Notes in Computer Science*. Springer.

Bavelas, J. B., & Gerwing, J. (2011). The listener as addressee in face-to-face dialogue. *International Journal of Listening, 255*(3), 178-198.

Beskow, J., & Al Moubayed, S. (2010). Perception of Gaze Direction in 2D and 3D Facial Projections. In *The ACM / SSPNET 2nd International Symposium on Facial Analysis and Animation*. Edinburgh, UK.

Edlund, J., Al Moubayed, S., & Beskow, J. (2011). The Mona Lisa Gaze Effect as an Objective Metric for Perceived Cospatiality. In Vilhjálmsson, H. H., Kopp, S., Marsella, S., & Thórisson, K. R. (Eds.), *Proc. of the Intelligent Virtual Agents 10th International Conference (IVA 2011)* (pp. 439-440). Reykjavík, Iceland: Springer.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., & Tapias, D. (Eds.), *Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (pp. 2992 - 2995). Valetta, Malta.

Kato, H., Takemoto, H., Nishimura, R., & Mokhtari, P. (2010a). Spatial acoustic cues for the auditory perception of speaker's facing direction. In *In Proc. of 20th International Congress on Acoustics, ICA 2010*. Sydney, Australia.

Kato, H., Takemoto, H., Nishimura, R., & Mokhtari, P. (2010b). On the human ability to auditorily perceive human speaker's facing angle. In *In Proc. of the 4th International Universal Communication Symposium (IUCS), 2010* (pp. 387 - 391). Beijing.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, 22-63.

Nakano, A. Y., Nakagawa, S., & Yamamoto, K. (2010). Auditory perception versus automatic estimation of location and orientation of an acoustic source in a real environment. *Acoustical Science and Technology, 31*(5), 309-319.

Nakano, A. Y., Yamamoto, K., & Nakagawa, S. (2008). Auditory perception of speaker's position, distance and facing angle in a real enclosed environment. In *Proc. of Autumn Meeting of Acoustic Society of Japan* (pp. 525-526).

Neuhoff, J. G., Rodstrom, M-A., & Vaidya, T. (2001). The audible facing angle. *Acoustics Research Letters Online, 2*(4), 109-114.

Neuhoff, J. G. (2001). Perceiving acoustic source orientation in three-dimentional space. In *Proc. of the International Conference on Auditory Display*. Espoo, Finland.

Neuhoff, J. G. (2003). Twist and shout: audible facing angles and dynamic rotation. *Ecological Psychology, 15*(4), 335-351.