On the effect of the acoustic environment on the accuracy of perception of speaker orientation from auditory 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

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. The feature most often implicated for detection of sound source orientation is the inter-aural level difference - a feature which it is assumed is more easily exploited in anechoic chambers than in everyday surroundings. We expand here on our previous studies and compare detection of speaker orientation within and outside of the anechoic chamber. Our results show that listeners find the task easier, rather than harder, in everyday surroundings, which suggests that inter-aural level differences is not the only feature at play.

Index Terms: turn-taking, head pose, gaze, acoustic directionality

1. Introduction

These days, we frequently study dialogue within the situation in which the dialogue takes place: 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. And we increasingly study the most primary and original form of spoken dialogue: face-to-face interaction. Perhaps the most central visual features of face-to-face interaction are gaze and head pose shifts. With the steadily increasing interest in embodied systems, they are becoming equally important for spoken dialogue systems, especially since, 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 previously demonstrated that a listener can perceive a speaker's facing angle under normal conversational circumstances to a surprising extent [1]. And to the extent that human speakers' facing angles are important, the auditory perception of a speaker's facing angle is important as well.

Our previous study took place in a recreational area at an office. In this study, we extend this to also include an anechoic chamber and a noisy bar. We add the noisy bar to verify an informal finding from mock studies several years ago: that auditory perception of a speaker's facing angle is possible under almost any acoustic circumstances. We add the anechoic chamber in part for comparison, but also to challenge the assumption that auditory perception of facing angle is chiefly a function of that inter-aural level differences (ILD). If this assumption holds, we would expect precision of auditory perception of a speaker's facing angle to be considerably higher in an anechoic chamber, since this setting provides ideal circumstances to perceive ILD. From our preliminary experiments, however, we suspect that more is involved - perhaps a dynamic modelling and recognition of the acoustic environment takes place in the listener. In this case, we would expect better results in familiar and typical everyday locations.

2. Background and related work

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. [2] 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). [3] 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 [2], the JND varies with the distance to the loudspeaker and with the facing angle itself. The work is brought together and discussed in [4], 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. [5] and [6] 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, [7] and [8] 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 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. [9; 10]). 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 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. [8], 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 coherent visible and audible cues to facing direction as well as to perceive and interpret the user's facing direction.

3. Method

The studies published to date were all performed in studios or rooms designed to minimize or normalize echoes, we decided against this. As a step in our current focus on co-presence [11] and situated, embodied conversational partners [12], we choose to stress real everyday environments, sacrificing control for ecological validity.

3.1. The subject/target experimental paradigm

We employed a generalized and adapted version of the subject/target paradigm first used in [13]. A group of 5 subjects were placed in a semi-circle, so that they were all watching the same point at their centre. All stimuli are presented from this point. All subjects double as targets for the directional stimuli (hence the *subject/target paradigm*). During the experiment, directional stimuli were aimed at each of the subjects. The order was varied systematically, and the number of stimuli was such that each subject was targeted twice in one set of stimuli. A set of stimuli, then, contains a 2*5 stimuli. Once one set was completed, the subjects shifted their positions by one step and the process of presenting a set of 10 stimuli was repeated. The rotation was repeated 5 times, until each subject had been in each position once, making the total number of stimuli presented in an experiment 50.

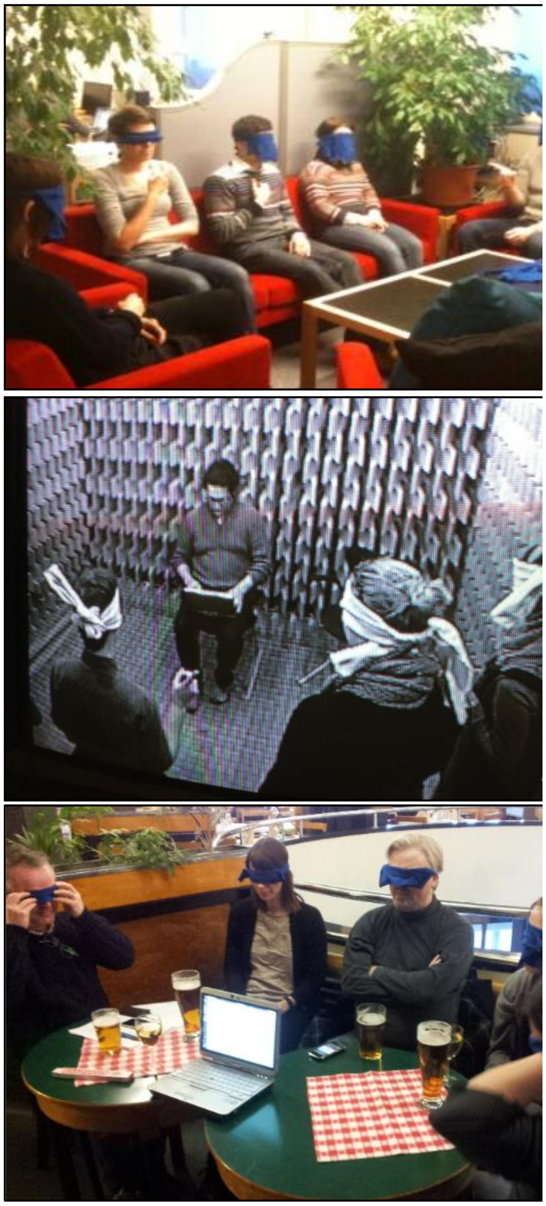

Figure 1: The three settings. The beers seen in the bar condition are for illustration and were not consumed during the experiment.

3.2. Settings

The main motivation for the experiment was to test the subjects' ability to perceive acoustic (speech) directionality in normal, everyday conditions. For this reason, the first setting (Recreational) was an existing recreational sofa group in busy office surroundings, and no attempts 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 the top of Figure 1. The second setting (ANECHOIC) was an anechoic chamber at the Linguistics department at Stockholm University, as seen in the middle panel of Figure 1.The third setting (BAR) was a busy bar housed in the rotund waiting building of a train station, with loud noises, much side talk, and echoes from irregularly shaped concrete walls, as seen in the bottom panel of the same figure.

A between-group design was employed. For the RECREATIONAL and ANECHOIC settings, two experiments were made, one with and one without intermittent feedback. We found no differences for these conditions [1], and due to limited time and access to subjects, the distinction was abolished for the BAR setting.

3.3. Subjects

The settings were tested in a between-group design with groups of five participants. The subjects were students and university employees. 11 of the subjects were female and 14 were male. All reported having normal hearing on both ears.

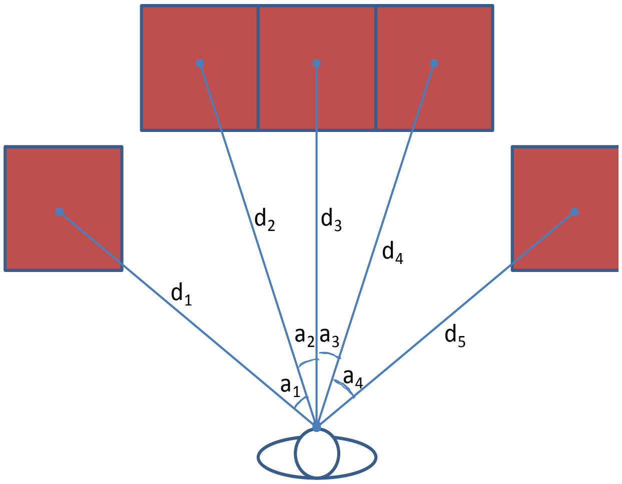

Figure 2: Schematic view of the experimental setup

3.4. Spatial layout

A result of two of the setting being actual and unaltered environments where people socialize and interact was that the distance to the nominal "centre" from which stimuli were presented was not identical for all seats and settings. The distances to the centre and the angles between each subject's position and that of the centre, as visualized in Figure 2, are presented in Table 1.

Table 1: Distances d_1 - d_5 (in cm) and angles a_1 - a_4 (in degrees) for each experiment setting.

	a_1	a_2	a_3	a_4	d_1	d_2	d_3	d_4	d_5
Anechoic	17	17	17	17	200	200	200	200	200
Recreational	28	18	18	28	160	190	180	190	160
Bar	36	29	29	36	120	160	140	160	120

3.5. Stimuli and responses

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 time a stimulus had been presented, each subject was asked to point out the intended target in such a manner that the other subjects could not take note of it. The result was 250 data points in one experiment.

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).

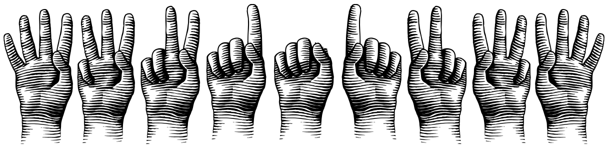

Figure 3: Signs used to indicate target position

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.

4. Results

Combined over all three settings, the subjects got the target exactly right in 259 out of 500 cases, or 44 % of the time, where random choice yields a 20 % baseline. A chi-square test on the contingency table yields shows that the result deviates significantly from a random choice ($\chi^2(1, N=16)=824$, $p<0.0001$), and the same test on contingency tables of the individual settings Recreational, Anechoic and Bar yield $\chi^2(1, N=16)=489$, $p<0.0001$; $\chi^2(1, N=16)=177$, $p<0.0001$; and $\chi^2(1, N=16)=177$, $p<0.0001$, respectively.

The confusion matrix for all data is shown in Table 2.

Table 2. Confusion matrix for all subjects and settings

		Estimated target position					
		1	**2**	**3**	**4**	**5**	**Total**
Target position	**1**	**125**	78	32	11	4	**250**
	2	35	**104**	86	23	2	**250**
	3	29	48	**98**	57	18	**250**
	4	8	23	89	**83**	47	**250**
	5	2	7	27	77	**137**	**250**
Total		**199**	**260**	**332**	**251**	**208**	**2500**

The average errors (Error) for the three settings were 13 degrees for Recreational, 16 degrees for Anechoic, and 18 degrees for Bar. We ran an ANOVA with the Error as the dependent variable, and Setting (Recreational, Anechoic, Bar) and Gender (Female, Male) as fixed independent variables. A Bonferroni Post Hoc test was used to test for differences between the different levels of Setting. These analyses showed a significant effect of Setting *F(2, 1244)=7.4; p<.05,* but neither Gender nor the interaction between Gender and Setting reached significance. The Bonferroni Post Hoc test showed that the Recreational setting had a significantly lower error *(p<.05)* than both Anechoic (-3.1 degrees) and Bar (-4.5 degrees), while there was no significant difference between Anechoic and Bar.

5. Discussion and future work

Our study verifies the finding of [2] and others: that listeners are quite good at distinguishing between different facing angles in a speaker. We further find that this is true not only in anechoic chambers and emptied out, quiet rooms, but also under conditions in which conversations normally occur - in furnished, asymmetric rooms with background noise and people passing by, and even in extremely noisy bars with under continuous bombardment of a multitude of sounds and echoes. In fact, we find that performance in the anechoic chamber is worse than in the relatively noisy recreational environment, and not significantly different from that of the very noisy bar.

The finding that speaker facing angle can be perceived by people in very real everyday environments - the environments in which conversations usually take place - 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. The redundancy is due to the much stronger visual cues that are often present - we can see the speaker's head orientation, but if our vision is somehow diminished, obscured, or otherwise out of order, we have access to acoustic cues that can help. It is also possible that conflicting acoustic and visual cues may increase cognitive load.

We argue that modelling the acoustic properties of speakers' position and orientation is an important step towards achieving a realistic model of situated interaction. Embodied spoken dialogue systems that aim for humanlike behaviour should present coherent and believable visual and acoustic cues. In the case of physical avatars and robots that use head pose and gaze for communicative purposes (e.g. 14; 15; 12), this could be done by embedding directional loudspeakers into their heads.

Our results also suggests that listeners use more than ILD to judge the facing angle of a speaker, and support the notion that they maintain an model of their acoustic environment into which they fit acoustic stimuli. The errors in the anechoic chamber are distributed symmetrically relative those in the other conditions, where certain errors are more common than others. In particular, subjects found it easy to pinpoint the rightmost speaking direction in the recreational area. One possible reason for this is that in that direction, there is a large window just behind the target and perpendicular to the speaker. The reflections from this window may well help subjects recognize that particular direction.

6. Acknowledgements

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

7. References

[1] Edlund, J., Heldner, M., & Gustafson, J. (2012). Who am I speaking at? - perceiving the head orientation of speakers from acoustic cues alone. In Proc. of LREC Workshop on Multimodal Corpora 2012. Istanbul, Turkey.

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

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

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

[5] Kato, H., Takemoto, H., Nishimura, R., & Mokhtari, P. (2010). 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.

[6] Kato, H., Takemoto, H., Nishimura, R., & Mokhtari, P. (2010). 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.

[7] 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).

[8] 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.

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

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

[11] Edlund, J., Al Moubayed, S., & Beskow, J. (2011). The Mona Lisa Gaze Effect as an Objective Metric for Perceived Cospatiality. In Proc. of the Intelligent Virtual Agents 10th International Conference (IVA 2011) (pp. 439-440). Reykjavík, Iceland: Springer.

[12] Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. In Esposito, A., Esposito, A., Vinciarelli, A., Hoffmann, R., & C. Müller, V. (Eds.), Cognitive Behavioural Systems. Lecture Notes in Computer Science. Springer.

[13] 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.

[14] Delaunay, F., de Greeff, J., & Belpaeme, T. (2010). A study of a retro-projected robotic face and its effectiveness for gaze reading by humans. In Procs of the 5th ACM/IEEE international conference on Human-robot interaction, ACM (pp. 39-44).

Kuratate, T., Matsusaka, Y., Pierce, B., & Cheng, G. (2011). Mask-bot: a life-size robot head using talking head animation for human-robot communication. In Proc. of the 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids) (pp. 99-104).