PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Audio-visual scene analysis in reverberant multi-talker environments

Axel Ahrens; Kasper Duemose Lund; Torsten Dau

Hearing Systems Section, Department of Health Technology, Technical University of Denmark, Denmark

ABSTRACT

Normal-hearing subjects are able to localize and identify sound sources in reverberant multi-talker environments. It has been shown previously that subjects can accurately analyse reverberant multi-talker scenes with up to four simultaneous talkers. While multi-talker scene perception has mainly been investigated regarding only auditory information, visual information might influence the subjects' perception. In the present study, audio-visual scenes varying between two and ten talkers were considered. The acoustic information was provided using a spherical loudspeaker array and visual information was provided using head-tracked virtual reality glasses. The visual information represented 21 possible static talker locations and the subjects were asked to identify the content of the talkers and their specific locations. For the identification of talkers, the subjects were asked to label visual locations with headlines from the talkers' speech topics. The subjects were able to accurately analyse scenes containing up to six talkers. When more talkers were presented in the scene, the azimuth localization accuracy decreased, whereas distance perception accuracy was not found to vary with the number of talkers. This new audio-visual scene analysis method might be a valuable tool to test speech perception in more realistic environments than those tested in previous investigations.

Keywords: Auditory Scene Analysis, Speech Perception, Virtual Reality

1. INTRODUCTION

The human auditory system is able to parse complex auditory scenes, also known as a "cocktail-party scenario" (see (1) for a review). The term cocktail-party scenario was coined by Colin Cherry (2) who investigated the perception of two simultaneous talkers presented over headphones. The speech was presented anechoically and either diotically or dichotically. Many studies investigated auditory perception in more complex listening scenarios with multiple talkers spatially distributed and including room reverberation (1). However, it has remained unclear how the human auditory system processes such scenarios.

Single speech or speech-like sound sources can be localized by human listeners with a resolution of up to a few degrees (e.g. 3). However, in the case of interfering stimuli presented and at negative signal-to-noise ratios, the localization accuracy has been shown to decrease (4–7). Kopco et al. (7) investigated the ability to identify and localize a target word spoken by a female talker out of four male interfering voices. They showed that the presence of the interferers disrupted speech localization accuracy relative to the condition word localization in quiet. However, the intelligibility of the target word was not considered in this study.

Hawley et al. (8) studied both localization accuracy and intelligibility of a target sentence in the presence of interfering speech. Up to three interfering talkers were included and the localization accuracy was generally found to be high and not correlated to the number of interfering talkers. The speech intelligibility was found to decrease with increasing number of interfering talkers, while the target-to-masker ratio was kept constant.

Most studies considered short speech stimuli, such as words or sentences. However, in realistic cocktail-party scenarios, the auditory system has generally access to longer intervals of speech. Weller et al. (9) used up to six scripted monologues, which were presented simultaneously in a loudspeaker-based virtual sound environment from locations placed around the listener. The listeners were asked to localize and identify the gender of the talkers in a top-down view representation of the listening environment. The results showed that listeners can reliably analyze a scene of up to four listeners. However, the authors noted that the analysis method was limited as the label for each source was binary (male/female) such that ambiguities could occur in the analysis.

In the current study, the ability of normal-hearing listeners to analyze a virtual audio-visual scene was investigated, containing a varying amount of talkers. The number of talkers in the scene was varied and the task was to specify the location in azimuth and distance and to identify the topic (content) of each source. The virtual audio-visual scenes were reproduced using a loudspeaker-based virtual sound environment and virtual reality glasses, allowing for an egocentric response.

2. METHODS

2.1 Participants

Six young, normal-hearing participants (5 female, 1 male) took part in the experiment. The participants were between 21 and 26 years with an average age of 23 years and were financially compensated on an hourly basis. All participants provided informed consent and all experiments were approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391).

2.2 Acoustic Reproduction

Ten stories were recorded in Danish language for this experiment. Each of the stories was read by 10 speakers (5 female, 5 male). The speech stimuli were recorded in a sound-proof listening booth using a Neumann TLM 102 (Neumann GmbH, Berlin, Germany) microphone. The text of the stories was shown on the head-mounted display of an HTC Vive Pro (HTC Corporation, New Taipei City, Taiwan) to eliminate acoustic reflections from a computer screen or noise from a paper. The speakers could re-read sections, but no particular focus was put on accuracy. The stories were between 74 and 120 s long and had an average duration of 93 s.

The reproduction of the stories was done in an anechoic room with a 64-channel spherical loudspeaker array with a radius of 2.4 m (10). The KEF LS50 (GP Acoustics Ltd., Maidstone, UK) loudspeakers were driven by three sonible d:24 (sonible GmbH, Graz, Austria) amplifiers.

The anechoic recordings of the stories were spatialized using the room acoustics simulation software Odeon (Odeon AS, Kgs. Lyngby, Denmark) and a nearest loudspeaker mapping method (NLM) from the loudspeaker auralization toolbox (LoRA, (11)). The NLM maps the direct sound as well as the early reflections to the geometrically closest loudspeaker. Late reflections are reproduced with energy envelopes represented in 1st order ambisonics and multiplied with uncorrelated noise for each loudspeaker (11).

2.3 Visual Reproduction

The virtual visual environment was reproduced using an HTC Vive (HTC Corporation, New Taipei City, Taiwan) virtual reality system. The head-mounted display, the hand-held controller and three additional Vive Trackers were tracked using the infrared ray tracking system. The Vive Trackers were used for a spatial calibration system as described in (12). Unity 3D with the SteamVR plugin were used to create and present the visual virtual environment.

2.4 Audio-Visual Environment

The virtual environment was a rectangular room of 9 by 12 meter and a height of 2.8 meter. Figure 1 shows the dimensions, the listener location and the possible source locations. The source locations were arranged semi-circularly around the listener between -90° (left) and 90° (right) in 30° steps. Three distances between the listener and the sources were included, 1.4m, 2.4m and 3.4m. On each possible source location, a static, semi-transparent avatar was visualized as shown in Figure 2. The participants could interact with the visual scene by pointing towards an avatar using a hand-held controller with a virtual laser pointer. A button on the hand-held controller could be pressed to choose an avatar, while another button changed the color of the laser. Additionally, a third button was assigned to choose the perceived distance.

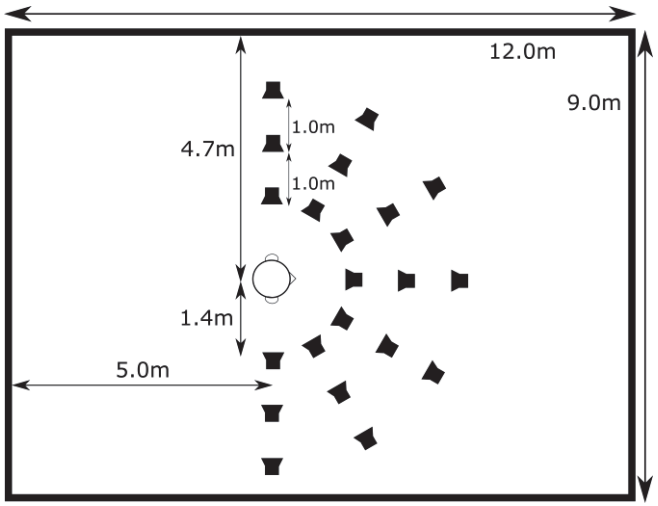

Figure 1 - Sketch of the virtual room with the listener location and the possible source locations. The height of the virtual room was 2.8 m.

2.5 Experimental Procedure

The experiment started with a familiarization phase, where subjects were asked to localize and identify a single source. In the familiarization phase, all ten stories were presented once to the listeners in random order. The talker and the location were drawn randomly on each trial.

Subjects were provided with a color-to-story map as shown on the back wall in Figure 2 (right panel). Identical colors were available for the laser pointer, enabling listeners to assign stories to locations. After each trial in the training phase, feedback of the correct location was provided to the listeners. The audio was played for 60 s, with no response time limit.

In the test phase of the experiment, the number of sources in the scene was varied between two and ten talkers. For each source, a unique talker, story and location were randomly chosen. No limitations regarding the distribution of the locations were considered. The audio was played for 120 s, without a time limit for the listeners' response. If a story was shorter than 120 s, it was re-started from the beginning. The subjects could indicate that they had finished analyzing the scene by pressing a button. Each scene complexity (two to ten sources) was tested three times with random source locations, stories and talkers. The order of the scene complexity conditions was randomized.

Figure 2 - Left: View on the virtual visual scene including the room and the possible source locations (semi-transparent avatars). Right: View from the participant's position. The colored icons indicate the topics of the stories, which correspond to the color of the laser pointer. The clock on the back wall indicates the remaining time in the current scene.

3. RESULTS

Figure 3 shows the correctly identified number of sources (panel A) and the number of perceived sources (panel B) in a scene as a function of the number of presented sources. The grey symbols show the individual results of the listeners and the black squares indicate the mean over all listeners. When up to six sources were presented in a scene, the listeners were, on average, able to accurately estimate the number of sources (panel A). For more than six sources, the accuracy decreased. Generally, the number of sources was underestimated when more than six sources were presented (panel B). Some subjects were found to identify the number of sources correctly until seven sources, while others had incorrect responses when four sources were presented simultaneously.

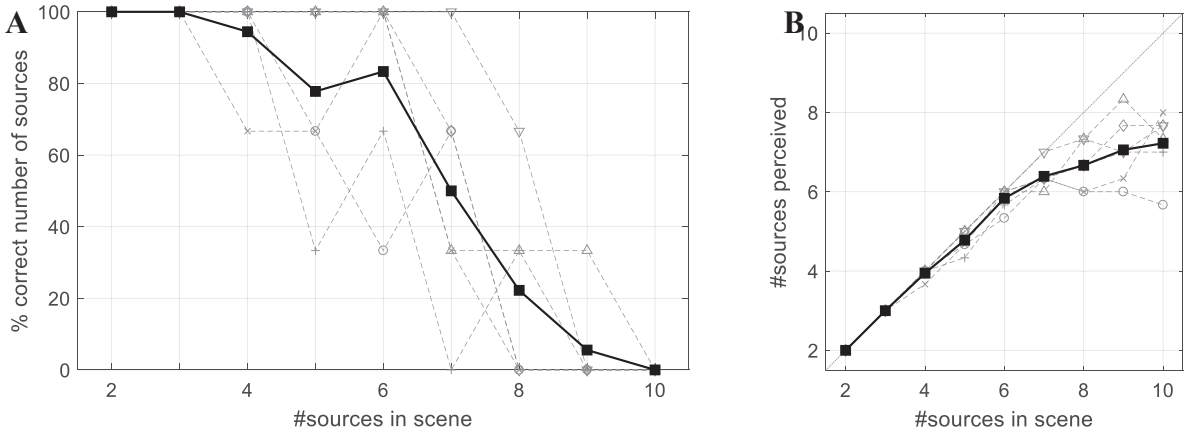

Figure 3 – Percent correct identified number of sources in a scene (A) and the number of perceived sources in a scene (B) as a function of the number of sources presented. The open symbols (grey lines) represent the individual listeners (averaged over the three repetitions) and the black squares represent the mean over the listeners.

Figure 4 shows the median time to complete a trial as a function of the number of sources in a scene. Even though the audio was played for only two minutes, the listeners took some extra time after the audio was finished when many sources were present. When only a few sources were presented in a scene, the time to complete the analysis was substantially lower.

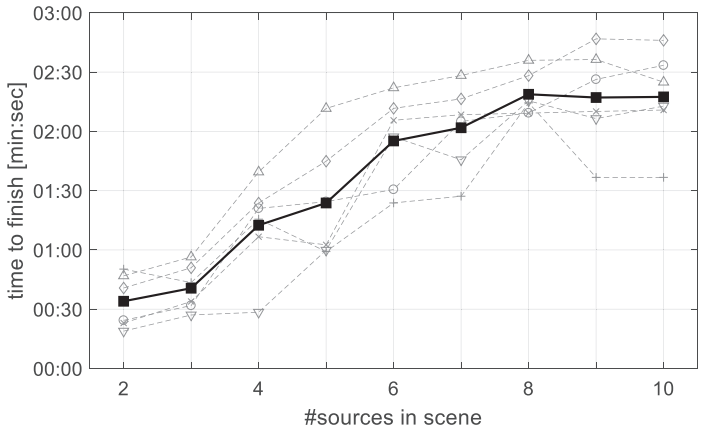

Figure 4 - The time to complete a single scene as a function of the number of sources. The open symbols (grey lines) represent the median over the three repetitions of the individual listeners and the black squares represent the mean over the listeners.

Figure 5 shows the spatial accuracy of the responses as a function of the number of sources in a scene. Panel A shows the root-mean square (RMS) azimuth error and panel B shows the RMS distance error. The error calculation only includes occurrences where a talker was perceived in the scene, thus, when the subjects wrongly estimated the number of talkers, only the responded talker was included in this analysis. The subjects correctly identified the azimuth direction for up to five sources and the error remained low also for six and seven sources. When more sources were presented in the scene, the error increased to up to 10°. The RMS distance error was, on average, found to be 0.4 m and roughly independent of the number of sources in the scene.

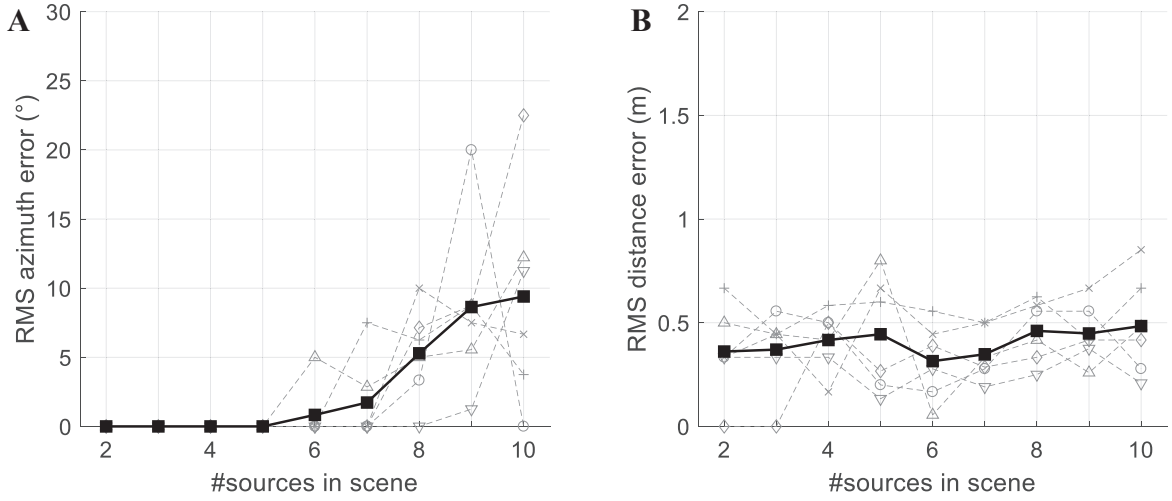

Figure 5 - Accuracy of the azimuth (A) and distance (B) perception as a function of number of sources. The open symbols (grey lines) represent the individual listeners, averaged over the three repetitions, and the black squares represent the mean over the listeners.

4. DISCUSSION

In the current study, the limit of accurately analyzing a scene was found to be at six talkers, which differs from the results obtained in Weller et al. (9) where a limit of four talkers was observed. Several differences between the two studies might have contributed to this difference. In the current study, the listeners heard 2 min. of audio, whereas Weller et al. (9) only played 45 s. Only the scenes with two and three talkers were finished within 45 seconds in the present study. In these conditions, the results of the two studies were in fact similar.

The RMS errors of both azimuth and distance localization were lower than in Weller et al. (9). This might be due to the lower number of possible azimuth directions and distance considered in the current study. Furthermore, in the current study a direct and egocentric response method (first-person view) was used, in contrast to the top-down representation of the scene in Weller et al. (9). It is possible that this translation from the first-person percept to the top-down response requires additional processing and thus reduces the accuracy (13,14).

In the current study an increased azimuth error was found for large number of sources. The increased azimuth error in these more complex scenes could possibly be explained by the task. The listeners were instructed to indicate the location of a perceived source even when it was unintelligible. Thus, there is a greater risk that a random story is indicated, which matches another story elsewhere in the scene. Here, the listeners were asked to prioritize the task of indicating the number of perceived sources over the task of accurately localizing the sources on the expense of a possible inflation of localization error in a scene with a large number of talkers. However, that would also have increased the distance perception error, which was not found to depend on the number of talkers in a scene. Furthermore, Weller et al. (9) specifically instructed their subjects to not guess if a talker was unintelligible and found generally higher errors than in the current study.

5. CONCLUSIONS

The ability of listeners to analyze audio-visual scenes with varying complexity was investigated. The complexity was varied by changing the number of simultaneous talkers. The audio-visual scenarios were reproduced using a loudspeaker-based virtual sound environment and head-tracked virtual reality glasses. The acoustic stimuli consisted of spatialized monologues and the visual stimulus of avatars without lip movements. It was found that subjects were able to accurately analyze such a virtual audio-visual scene in conditions with up to six talkers. When more talkers were presented, the detection accuracy of the number of sources decreased as did the azimuth localization accuracy.

The presented method allows for a more realistic testing than traditional speech intelligibility test paradigms as the task of analyzing a complex scene appears more realistic than repeating single words or sentences.

ACKNOWLEDGEMENTS

We would like to thank Marton Marschall, Valentina Zapata Rodriguez and Jakob Nygård Wincentz for their valuable feedback regarding the audio-visual environment.

REFERENCES

1. Bronkhorst AW. The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions. Acta Acust united with Acust [Internet]. 2000;86(1):117–28. Available from: http://link.springer.com/10.3758/s13414-015-0882-9
2. Cherry EC. Some Experiments on the Recognition of Speech, with One and with Two Ears. J Acoust Soc Am [Internet]. 1953 Sep;25(5):975–9. Available from: http://asa.scitation.org/doi/10.1121/1.1907229
3. Blauert J. Spatial hearing: the psychophysics of human sound localization. MIT Press; 1997.
4. Lorenzi C, Gatehouse S, Lever C. Sound localization in noise in normal-hearing listeners. J Acoust Soc Am [Internet]. 1999 Mar;105(3):1810–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10089604
5. Good MD, Gilkey RH. Sound localization in noise: The effect of signal-to-noise ratio. J Acoust Soc Am [Internet]. 1996 Feb;99(2):1108–17. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8609294
6. Abouchacra KS, Emanuel DC, Blood IM, Letowski TR. Spatial perception of speech in various signal to noise ratios. In: Ear and Hearing. 1998. p. 298–309.
7. Kopčo N, Best V, Carlile S. Speech localization in a multitalker mixture. J Acoust Soc Am [Internet]. 2010;127(3):1450–7. Available from: http://asa.scitation.org/doi/10.1121/1.3290996
8. Hawley ML, Litovsky RY, Colburn HS. Speech intelligibility and localization in a multi-source environment. J Acoust Soc Am [Internet]. 1999;105(6):3436–48. Available from: http://asa.scitation.org/doi/10.1121/1.424670
9. Weller T, Best V, Buchholz JM, Young T. A Method for Assessing Auditory Spatial Analysis in Reverberant Multitalker Environments. J Am Acad Audiol [Internet]. 2016;27(7):601–11. Available from: http://openurl.ingenta.com/content/xref?genre=article&issn=1050-0545&volume=27&issue=7&spage=601
10. Ahrens A, Marschall M, Dau T. Measuring and modeling speech intelligibility in real and

loudspeaker-based virtual sound environments. Hear Res [Internet]. 2019 Feb; Available from: https://linkinghub.elsevier.com/retrieve/pii/S0378595518305598

11. Favrot S, Buchholz JM. LoRA: A loudspeaker-based room auralization system. Acta Acust united with Acust [Internet]. 2010 Mar 1 [cited 2017 Oct 19];96(2):364–75. Available from: http://openurl.ingenta.com/content/xref?genre=article&issn=1610-1928&volume=96&issue=2&spage=364

12. Ahrens A, Lund KD, Marschall M, Dau T. Sound source localization with varying amount of visual information in virtual reality. Malmierca MS, editor. PLoS One [Internet]. 2019 Mar 29;14(3):e0214603. Available from: http://dx.plos.org/10.1371/journal.pone.0214603

13. Town SM, Brimijoin WO, Bizley JK. Egocentric and allocentric representations in auditory cortex. Ungerleider L, editor. PLOS Biol [Internet]. 2017 Jun 15;15(6):e2001878. Available from: https://dx.plos.org/10.1371/journal.pbio.2001878

14. Schlag J, Schlag-Rey M. Through the eye, slowly; Delays and localization errors in the visual system. Nat Rev Neurosci [Internet]. 2002 Mar;3(3):191–191. Available from: http://www.nature.com/articles/nrn750