A method for evaluating audio-visual scene analysis in multi-talker environments

KASPER D. LUND[1]*, AXEL AHRENS[1] AND TORSTEN DAU[1]

[1] *Hearing Systems Section, Department of Health Technology, Technical University of Denmark*

In cocktail-party environments, listeners are able to comprehend and localize multiple simultaneous talkers. With current virtual reality (VR) technology and virtual acoustics it has become possible to present an audio-visual cocktail-party in a controlled laboratory environment. A new continuous speech corpus with ten monologues from five female and five male talkers was designed and recorded. Each monologue contained a substantially different topic. Using an egocentric interaction method in VR, subjects were asked to label perceived talkers according to source position and content of speech, while varying the number of simultaneously presented talkers. With an increasing number of talkers, the subjects' accuracy in performing this task was found to decrease. When more than six talkers were in a scene, the number of talkers was underestimated and the azimuth localization error increased. With this method, a new approach is presented to gauge listeners' ability to analyze complex audio-visual scenes.

INTRODUCTION

Normal-hearing listeners are able to localize and understand multiple talkers in complex listening environments, also referred to as 'cocktail-party' scenarios (Bronkhorst, 2000). The ability of the auditory system to analyze such complex scenes is often referred to as "auditory scene analysis". Previous studies have employed signal patterns with varying degrees of spectral or temporal differences to investigate how the auditory system analyses scenes (Bregman, 1994). Other studies have used more speech-like stimuli to increase the ecological validity. However, these test paradigms might not reflect perception in more realistic complex acoustic scenes.

Kopčo *et al.* (2019) asked subjects to identify the location of a female talker in a mixture of male talkers and showed a reduction in localization accuracy relative to a condition without interferers. Weller *et al.* (2016) simulated a more realistic auditory scene with up to six simultaneous continuous speech sources in a reverberant room and asked subjects to identify the location and the gender of the talkers. The subjects were provided a top-down view of the room on a touchscreen. Weller *et al.* (2016) found that normal-hearing subjects were able to accurately analyze scenes with up to four talkers.

*Corresponding author: kdue@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

Even though the realism of the paradigms investigating multi-talker scene analysis has increased, some factors have not been considered. For example, most studies focused on audio-only settings, or used allocentric interfaces, where subjects do not have a first-person view of the scene. Thus, potential influences of visual cues and egocentric perception have not been considered.

With recent advances in virtual reality (VR) technology, it is possible to present visual content via a head-mounted display (HMD) in a controlled environment. In the present study, we propose a novel method for investigating complex scene analysis in a realistic audio-visual setup, by combining VR technology, virtual acoustics, and by utilizing an egocentric interface.

METHODS

Speech stimuli

To create the auditory scenes a speech material corpus was designed. The corpus was established by recording continuous Danish speech material, consisting of monologues on substantially different topics. Ten monologues with easy readability and unique terms and words to maximize their distinguishability were composed. Ten non-professional native Danish speakers (five females, five males) were individually recorded while reading each of the ten stories. The fundamental frequencies of the talkers ranged from 116 to 204 Hz. The talkers were recorded with a Neumann TLM 102 large diaphragm condenser microhpone (Neumann GmbH, Berlin, Germany) in a sound-proof listening booth. The text was presented on a virtual teleprompter on an HTC Vive Pro VR system (HTC Corporation, New Taipei City, Taiwan) to avoid the noise from paper or acoustic reflections from a computer screen. Using the VR controller, talkers could scroll through the text in their own pace. For optimal readability the virtual teleprompter was adjustable in distance (size) and height. Each monologue recording was equalized to the same root-mean-square level.

Acoustic setup

The experiment was conducted in an anechoic room containing a 64-channel loudspeaker array (see Ahrens *et al.*, 2019a, for details). The dry speech recordings were spatialized in a simulated reverberant room created with Odeon (Odeon A/S, Kgs. Lyngby, Denmark). The loudspeaker signals were generated employing a nearest loudspeaker mapping method using the loudspeaker auralization toolbox (LoRA Favrot and Buchholz, 2010). The room had an average reverberation time of $\sim$0.4 s. Fig. 1 (left panel) shows an overview of the room and the 21 possible talker positions. The positions were all at ear level and in the frontal hemisphere from -90° (left) to 90° (right), in 30° steps. Three distances were considered for all azimuth directions, 1.4m, 2.4m and 3.4m, where the 2.4m distance coincided with the radius of the loudspeaker array.

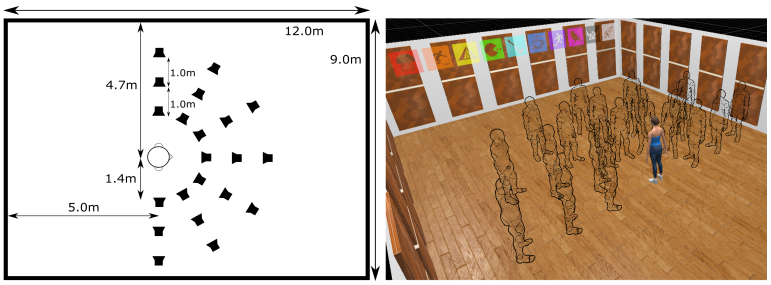

Fig. 1: (Left) Overview of the acoustic scene setup. The room was 2.8m high. The 21 simulated talker positions are indicated by loudspeaker symbols. (Right) Overview of the visual scene setup visually. Semi-transparent humanoid bodies were positioned at locations corresponding to the acoustic source positions.

Visual setup

The visual scene was presented on a HMD (HTC Vive system, HTC Corporation, New Taipei City, Taiwan). Fig. 1 (right panel) shows the virtual visual scene. It contained a room that visually matched the simulated acoustic room in terms of size and surface materials as well as 21 semi-transparent unisex humanoid bodies that were displayed at positions corresponding to the acoustic source positions. At the back wall of the the virtual room, a list of coloured icons was shown, representing the topics of all stories.

The visual scene was rendered using Unity software (Unity Technologies, San Francisco, California, USA) with the SteamVR plugin (Valve Corporation, Bellevue, Washington, USA). To ensure the spatial alignment between the acoustic and the visual scenes, a calibration method using VR trackers was employed (Ahrens *et al.*, 2019b).

Subjects and procedure

Six young (24,3 years old on average), self-reported normal-hearing, native Danish speaking subjects participated in the experiment. Prior to their participation all subjects gave their written consent to the ethics agreement approved by the Science-Ethics Committee for the Capital Region of Denmark (reference H-16036391). Three repetitions of scenes for each number of simultaneous talkers were run - 27 trials in total. Each subject completed the experiment within two hours and where allowed breaks after each trial.

On each trial, between two and ten talkers and stories were randomly chosen and simultaneously presented from random locations. Duplicates for talkers, stories and positions were not allowed. On each trial, the subjects were asked to identify the stories present in the scene and to change the color of the virtual laser pointer (as seen in Fig. 2) with a button press on a handheld VR controller to match the color

of an icon representing the perceived story. Another button was assigned to change the distance of the laser-pointer to mark sources at different distances. After choosing the color and the distance of the laser-pointer, the subjects could label the perceived talker location by choosing an avatar. After the selection, the avatar changed the color according to the color of the icon/laser-pointer. The audio was presented for 120 s. The time for the subjects' responses was not restricted, but finalized with a button press on the controller. Each acoustic talker was presented at a sound pressure level (SPL) of 55 dB and no feedback was provided to the subjects.

Before the test session, each subject participated in a familiarization session. Each of the ten stories were separately presented once to the subject. A random talker and location was assigned in each trial and the subject was asked to do the task as described above. The audio signals were presented for up to 60 s. After the response, feedback was provided to the subjects by indicating the correct story and position of the talker.

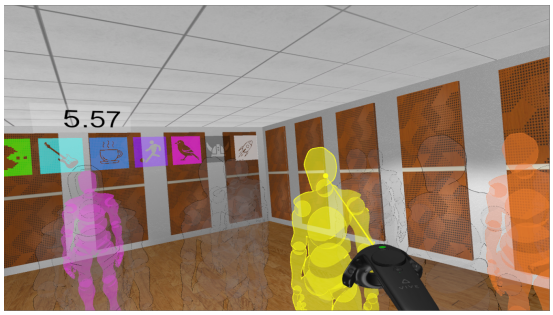

Fig. 2: The visual scene as seen from the point-of-view of the listener. Using the virtual laser pointer subjects' task was to analyze the acoustic scenes and label the perceived positions of a talker according with the perceived story.

RESULTS AND DISCUSSION

Fig. 3 shows how often each talker (top panel) and each story (bottom panel) was correctly identified. This measure evaluates the overall identification difficulty of talkers and stories in the collected response data. The top panel is split into female ('f', left) and male ('m', right) talkers. The analysis of a linear model showed no significant difference in talker identification difficulty ($F(8,50) = 1.37, p = 0.23$). However, an average difference between male and female talker identification accuracy of 9.3%-points was found ($F(1,50) = 10.01, p = 0.0026$), indicating that it was more difficult to identify the male compared to the female talkers. Previous studies showed similar trends (Bradlow *et al.*, 1996). The bottom panel of Fig. 3 shows the story identification difficulty. The analysis of the linear mixed model showed no significant difference across the stories ($F(9,45) = 1.25, p = 0.29$).

Fig. 4 shows the number of perceived talkers as a function of the number of presented talkers in the scene. The black squares represent the mean across subjects and the grey lines show the individual subject data.

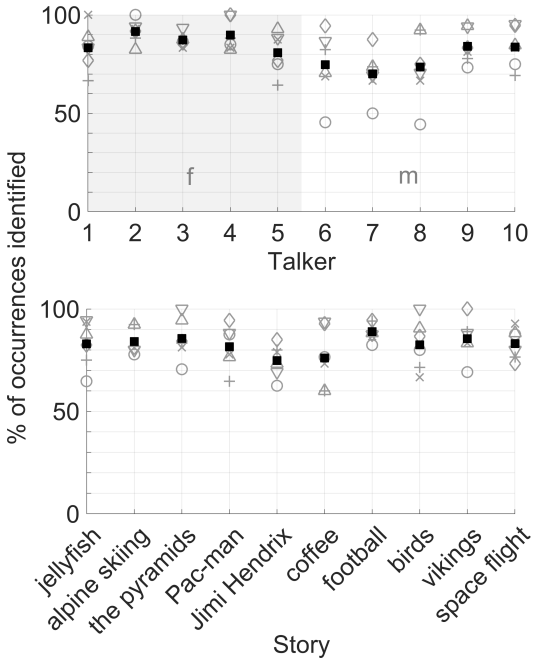

Fig. 3: Overall identification difficulty of talkers (top) and stories (bottom), as % of the occurrences identified. Open grey symbols represent individual subjects, averaged over three repetitions, and black squares represent the mean over subjects. Genders are indicated for talkers ('m' or 'f').

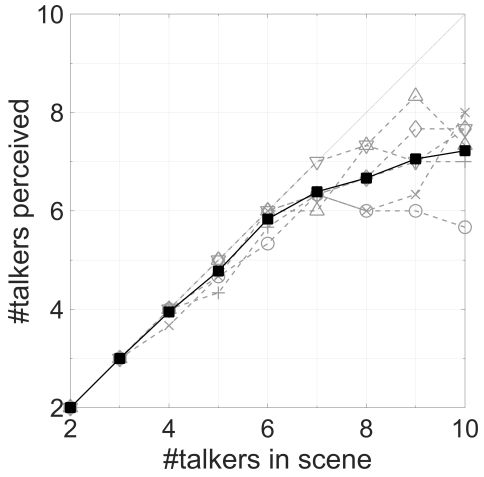

Fig. 4: Number of perceived talkers over number of presented talkers in the scene. Open grey symbols represent individual subjects, averaged over three repetitions, and black squares represent the mean over subjects.

For scenarios with up to six simultaneous talkers, the majority of the subjects were able to correctly identify the number of talkers. In scenarios with more than six

simultaneous talkers, this ability decreased gradually and, on average, the subjects underestimated the number of talkers in the scene.

Compared to the results from Weller *et al.* (2016), the number of correctly identified simultaneous talkers was higher in the present study. While the task in Weller *et al.* (2016) was similar, they only presented auditory stimuli (i.e. no visual information), used an allocentric interface for the response collection, and subjects were only given 45 s response time. Whether the observed larger number of identified talkers resulted from the visual gain, the egocentric interface or the increased time limit, still needs to be clarified. In Weller *et al.* (2016) the subjects needed to judge the gender of the talkers, while in the current study the content of the story needed to be identified. The identification of the content is likely to be more difficult and is expected to reduce the number of correctly identified talkers, which has not been observed in the current study.

Fig. 5 shows the localization accuracy of the sources as a function of the number of simultaneous talkers. The left panel shows the root mean squared (RMS) error in azimuth and the right panel shows the RMS error in distance. Individual subject responses are indicated by the grey symbols whereas the average results across subjects are shown as black squares. For up to five simultaneous talkers, all talker azimuth positions were correctly identified. For up to seven simultaneous talkers, the error did not significantly increase ($p < 0.0001$) as indicated by the analysis of a linear mixed model. For eight and more talkers, the error increased gradually. The distance error (right panel) was found to be independent of the number of the talkers ($F(8,148) = 0.79, p = 0.62$). The average RMS distance error was about 0.57m and thus below the chance level of 1.15m. The chance level was calculated as the RMS across all possible errors at all three distances.

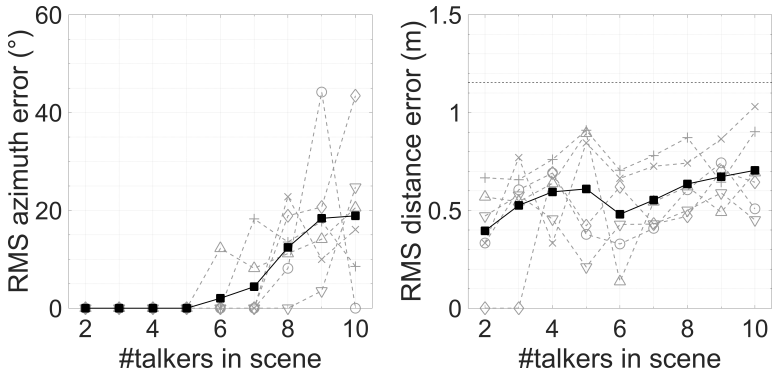

Fig. 5: Localization accuracy, in azimuth (left) and distance (right) RMS errors over number of talkers in scene. Open grey symbols represent individual subjects, averaged over three repetitions, and black squares represent the mean over subjects. The dotted, horizontal line represents the chance level.

Compared to Weller *et al.* (2016), the present study showed a higher localization accuracy, potentially resulting from the smaller spatial range of the response options. The additional visual information and/or the egocentric interface might also have improved the localization accuracy.

Fig. 6 shows the story identification ability of the subjects with respect to the number of simultaneous talkers. The story identification ability is represented as the percentage of scenarios where all stories were recognized. The analysis of a linear mixed model showed a significant effect of the number of talkers ($F(8,148) = 31.4, p < 0.0001$). The ability to identify the correct story decreased gradually with an increasing number of simultaneous talkers. On average, the subjects could identify stories correctly for up to five simultaneous talkers, whereas for eight or more talkers, none of the subjects were able to analyze any scene correctly. Thus, while the subjects were able to accurately identify the number of talkers up to six talkers, the speech recognition ability was only accurate up to five talkers.

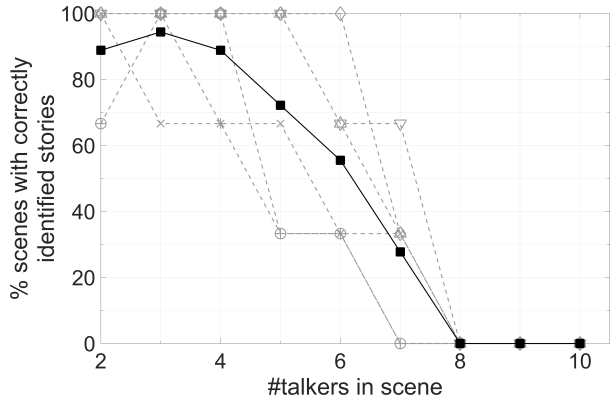

Fig. 6: Percentage of scenes analyzed correctly according to presented stories over number of talkers in scene. Open grey symbols represent individual subjects, averaged over three repetitions, and black squares represent the mean over subjects.

SUMMARY AND CONCLUSION

In the current study, a novel method for evaluating a subject's auditory scene analysis ability in realistic multi-talker scenes was proposed. The method allows to measure sound source identification and localization perception in an audio-visual environment using a loudspeaker-based virtual sound environment and a virtual reality headset. Compared to traditional sentence-based audio-only approaches, this method allows for testing in a more realistic environment and could possibly be used as a tool to evaluate hearing instruments and algorithms.

It was shown that subjects were able to identify the number of talkers in scenes with up to six simultaneous talkers. Furthermore, the localization ability was found to remain

unaffected for scenes with up to seven simultaneous talkers, while the perception of distance did not depend on the number of simultaneous talkers in the scene. The speech recognition ability was found to be worse than the identification of the number of simultaneous talkers.

The VR-based audio-visual method presented in the current study results in improved response accuracy and talker number identification ability compared to previous studies. Future investigations could address how different listening conditions additionally affect motion behavior, such as head rotation and eye gaze.

ACKNOWLEDGEMENTS

The authors would like to thank Marton Marschall, Jakob Nygård Wincentz and Valentina Zapata Rodriguez for feedback during the development of the simulated environments. Furthermore, we would like to thank the talkers for letting us record their voices.

REFERENCES

Ahrens, A., Marschall, M., and Dau, T. (**2019**). "Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments," Hearing Res., **377**, 307-317. doi: 10.1016/j.heares.2019.02.003.

Ahrens, A., Lund, K.D., Marschall, M., and Dau, T. (**2019**). "Sound source localization with varying amount of visual information in virtual reality," PLOS ONE, **14**(3), e0214603. doi: 10.1371/journal.pone.0214603.

Bradlow, A.R., Torretta G.M., and Pisoni, D.B. (**1996**). "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun., **20**(3-4), 255-27. doi: 10.1016/S0167-6393(96)00063-5.

Bregman, A.S. (**1994**). "Auditory Scene Analysis: The Perceptual Organization of Sound," MIT Press. doi: 10.1121/1.408434.

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acust united Ac, **86**(1), 117-128.

Favrot, S. and Buchholz, J.M. (**2010**). "LoRA: A loudspeaker-based room auralization system," Acta Acust united Ac, **96**(2), 364-375. doi: 10.3813/AAA.918285.

Kopčo, N., Best, V., and Carlile, S. (**2010**). "Speech localization in a multitalker mixture," J. Acoust. Soc. Am., **127**(3), 1450-1457. doi: 10.1121/1.3290996.

Weller, T., Best, V., Buchholz, J. M., and Young, T. (**2016**). "A method for assessing auditory spatial analysis in reverberant multitalker environments," J. Am. Acad. Audiol., **27**(7), 601-611. doi: 10.3766/jaaa.15109.