Proceedings of the 2025 SICE Festival with Annual Conference
September 9-12, 2025, Chiang Mai, Thailand

Development of a Spatial Audio Feedback System for Dynamic Object Localisation in the Visually Impaired

Hiroto Obu[1†], Ferina Ayu Pusparani[1],Wen Liang Yeoh[1], Fukuda Osamu[1]

[1]Department of Science and Engineering, Saga University, Saga City, Japan.
(E-mail: oobuhiroto@gmail.com, ferinaayu.01@gmail.com, {wlyeoh, fukudao}@cc.saga-u.ac.jp)

Abstract: One of the main challenges faced by individuals with visual impairments is understanding their surroundings due to the lack of visual information. Recent advancements in artificial intelligence (AI) have enabled applications that can automatically describe environments through voice feedback. However, these language-based systems often struggle to convey information about crowded and dynamic environments, such as train stations or busy intersections, making it difficult for users to track multiple moving elements. To address this, we propose the use of spatial audio, leveraging human sound localization capabilities to complement linguistic descriptions and improve spatial awareness. We developed a prototype system and investigated how two factors—the type of sound source (continuous vs. intermittent) and the distance attenuation model (linear vs. exponential)—affect localisation performance for both stationary and moving objects.

Keywords: Human Interfaces,Human-Machine Systems

1. INTRODUCTION

1.1. Background

It is estimated that approximately 43 million people worldwide are blind, and about 295 million live with moderate to severe visual impairment [1]. One of the main challenges faced by individuals with visual impairments is the difficulty in understanding their surroundings due to their inability to obtain information visually.

To navigate their environment, visually impaired individuals often rely on white canes, guide dogs, or assistance from other people. While white canes are useful for detecting obstacles, they require physical contact with objects and can only provide information about a single point in space at any given moment. This limits its range and makes it impractical for use to keep track of moving objects. Guide dogs, on the other hand, offer a higher level of support but come with significant time and financial costs for training and maintenance, making them difficult to adopt widely.

In recent years, we have seen rapid advancements in computing technologies, particularly in artificial intelligence (AI) for object and scene recognition, and in generating descriptions to explain images. In addition to traditional assistive devices, these emerging technologies are increasingly being used by visually impaired individuals to support their daily activities. For example, smartphone applications and smart glasses equipped with AI can recognise the surrounding environment using cameras and convey this information to the user through voice feedback [2], [3].

Although incredibly helpful in enabling people with visual impairment to gain an understanding of their surroundings in relatively sparse and static environments, these language-based voice feedback systems can often struggle to keep up in crowded and dynamic environments. For example, in train stations or at a busy intersection where there may be many vehicles and people moving simultaneously, it is very difficult to convey this information linguistically in real-time [4], [5].

In this study, we propose that spatial audio can be used to complement these linguistic explanations by helping the user keep track of the various moving objects. By leveraging the human ability to localise sound, it is possible to convey the real-time position and movement of objects in three-dimensional space—similar to how visually impaired para-athletes track a ball using the sound of the bell inside it.

1.2. Related Works

In recent years, there have been several studies that attempted to use spatial audio to help people with visual impairments navigate their surroundings.

Yang *et al.* developed a system that detects and classifies objects using a smartphone, providing audio feedback based on the classification. They varied the frequency, binaural balance, and amplitude of the audio to help users construct cognitive maps of their surroundings [6]. Hu *et al.* developed a wearable system that utilised an RGB-D camera to acquire 3D spatial information of the environment and compared the performance of spatial audio with that of speech instructional feedback. They found that spatial audio reduced positioning error by 40% [7].

These studies have demonstrated the potential and effectiveness of spatial audio in helping individuals with visual impairments perceive their three-dimensional surroundings. However, many challenges remain in developing an effective spatial audio-based assistive device. For instance, how does the type of sound affect localisation accuracy? Moreover, most existing research has focused on static objects, while the goal of this study is to explore the tracking of moving objects. It also remains unclear whether there is a difference in localization performance between moving and static targets.

† Hiroto Obu is the presenter of this paper.

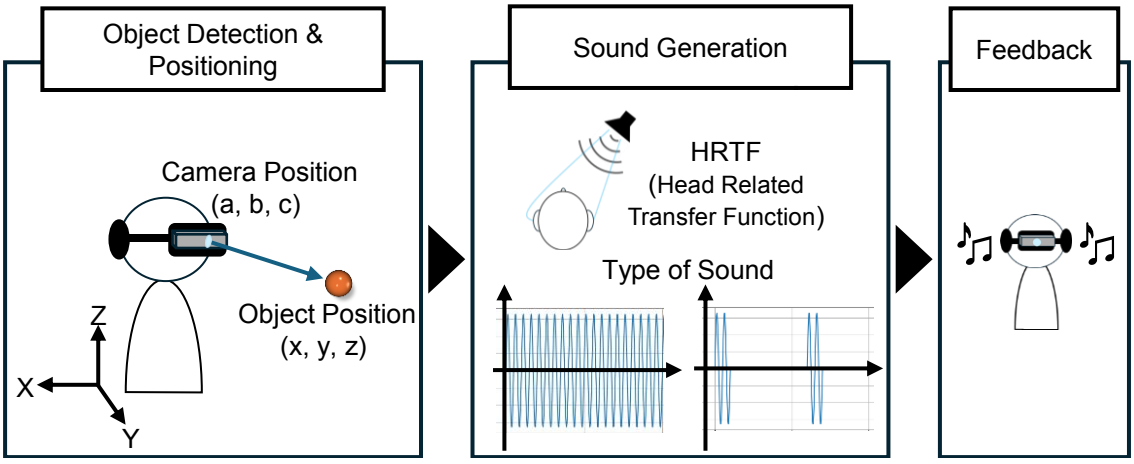


Fig. 1 Spatial Audio Feedback System Overview

1.3. Aim

As a first step, we aim to investigate the effect of the type of sound used and various spatial audio parameters on sound localisation capability for both static and moving objects. Specifically, whether the type of sound source used (continuous vs intermittent), and whether the distance model (linear or exponential) affects sound affects localisation performance in the case of a static object and a moving object. In order to do this, we developed a prototype system that tracks the position of a ball in real-time and provides spatial audio feedback based on the relative position of the ball to the user's head.

2. SYSTEM CONFIGURATION

An overview of the system developed in the study is shown in Fig. 1. The target object used in this study is an orange table tennis ball. This system consists of the following three elements: an RGB-D camera attached to the user's head, a computing device that detects that target object and calculates the appropriate spatial audio feedback, and an open ear headphone to feed back the generated binaural audio to the user. First, the three-dimensional location of the target object is detected using an RGB-D camera. Next, the system generates appropriate spatial audio feedback based on the obtained three-dimensional location information of the object. Finally, the generated sound is presented to the user via an open ear headphone.

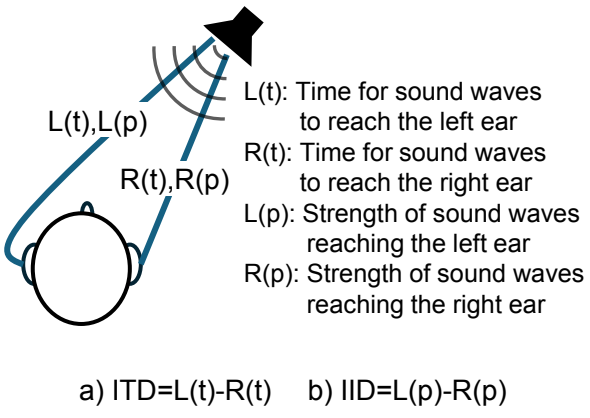


Fig. 2 Head-Related Transfer Function (HRTF)

2.1. Object Detection and Positioning

A wide angle RGB-D camera (Luxonis OAK-D W) is used to capture the 2D RGB scene as well as the depth map of the scene. Colour-based object detection is used to determine the pixel coordinate of the target object in the camera view. Using this pixel coordinate (u, v) and its corresponding depth value Z, the three-dimensional location of the ball relative to the user's head (X, Y, Z) can be computed as:

$$X = \frac{(u - c_x) \cdot Z}{f_x} \tag{1}$$

$$Y = \frac{(v - c_y) \cdot Z}{f_y} \tag{2}$$

$$Z = \text{depth at } (u, v) \tag{3}$$

where:

- f_x, f_y are the focal lengths in pixels,
- c_x, c_y are the coordinates of the principal point (optical center),
- Z is the depth value in meters.

2.2. Sound Generation

To simulate a three-dimensional sound environment, all generated audio will be using the same head-related transfer function (HRTF) described in Section 2.2.1. Section 2.2.2 describes the properties of the two types of sound sources to be investigated in this study. Finally, in Section 2.2.3, the two distance model to be compared is described.

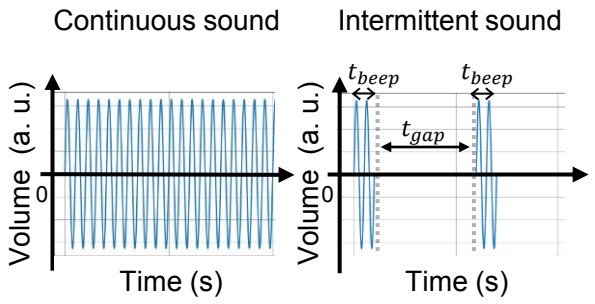
2.2.1. Head-Related Transfer Function

The Head-Related Transfer Function (HRTF) describes how a sound wave is filtered depending on its source location, based on anatomical features such as the

head shape, pinnae, and torso [8]. Specifically, it models the effects of the head, torso, and outer ear (auricle) on sound waves as they travel from free space to the eardrum. As described in Fig. 2, HRTF incorporates cues such as Interaural Time Difference (ITD) and Interaural Intensity Difference (IID) to reproduce the spatial location of sound sources in three dimensions. It is widely used in spatial audio technologies, including virtual reality applications, 3D audio in gaming, and stereoscopic sound systems in cinemas. In our system, we made use of a generic HRTF and the calculation was performed using the Synthizer library [9].

2.2.2. Continuous vs Intermittent

We hypothesise that an intermittent sound would allow us to better discern the changes in sound pressure level or volume due to the gaps between sound bursts. Therefore, we compare two types of sounds in this study, continuous and intermittent, as shown in Fig 3. Both sounds were pure sine waves of frequency 1 kHz. For the intermittent sound, gaps of 0.1 seconds separate sound bursts of period 0.05 seconds.



$t_{beep} = 0.05$ s, $t_{gap} = 0.1$ s, $f = 1$ kHz

Fig. 3 Type of sound (Continuous vs Intermittent)

2.2.3. Distance Model

In general, the sound pressure level tends to be related to the distance from the sound source [10], [11]. The louder the sound, the closer the object is perceived to be, and the quieter the object, the farther away it is perceived to be. In this study, we investigate the effects of two distance models, namely, linear and exponential. The two distance models are illustrated in Fig. 4.

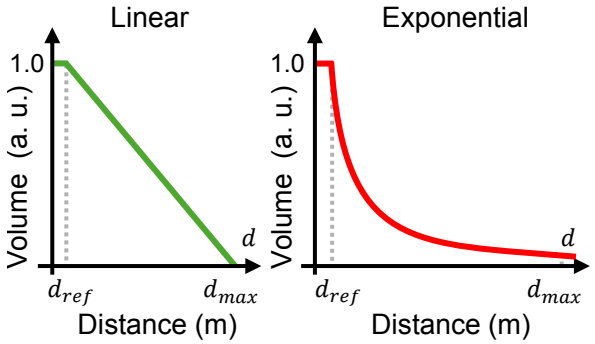
Firstly, the relative distance from the sound source to the camera is calculated using Eq. 4 where the (x, y, z) is the position of the sound source and (a, b, c) is the position of the camera. Because the position of the target object in this study is calculated relative to the camera in Section 2.1, (a, b, c) is 0 in our case.

$$d = \sqrt{(x-a)^2 + (y-b)^2 + (z-c)^2} \tag{4}$$

The change in sound pressure level or volume for the linear model $F(d)$ and the exponential mode $G(d)$ can then be calculated using Eq. 5 and Eq. 6. In both equations, r represents the sound attenuation rate, d_{ref} is the minimum distance below which the volume is maximum, and $d_{\max}$ is the maximum distance above which the volume is minimum. In this study, $d_{\max}$ of 4 m, d_{ref} of 0.3 m and r of 1 is used.

$$F(d) = 1 - r \cdot \frac{\text{clamp}(d, d_{\text{ref}}, d_{\max}) - d_{\text{ref}}}{d_{\max} - d_{\text{ref}}} \tag{5}$$

$$G(d) = \left(\frac{\max(d_{\text{ref}}, d)}{d_{\text{ref}}} \right)^{-r} \tag{6}$$



$d_{max} = 4$ m, $d_{ref} = 0.3$ m, $r = 1$

Fig. 4 Distance Model

2.3. Audio Feedback

The audio generated in Section 2.2 is fed back to the user in real-time using a wired open ear headphone (Nwm MWE001).

3. EXPERIMENT

The purpose of this experiment is to determine which type of sound and which distance model best supports accurate spatial recognition using only auditory cues, without any visual input. Two tasks were used to evaluate them: one involving a stationary object, and the other involves a moving object.

3.1. Experiment Task

3.1.1. Stationary Object

The experimental setup for this task is shown in Fig.5. The subjects sat in a chair and were instructed to search for a randomly placed table tennis ball within a designated area, relying solely on audio feedback provided by the system. The search area was defined based on the study by Yang *et al.* [12], with a radius of 60 cm—chosen as a comfortable range for arm movement while seated. Once the subject has determined the position of the ball, they were instructed to point to it using their index finger.

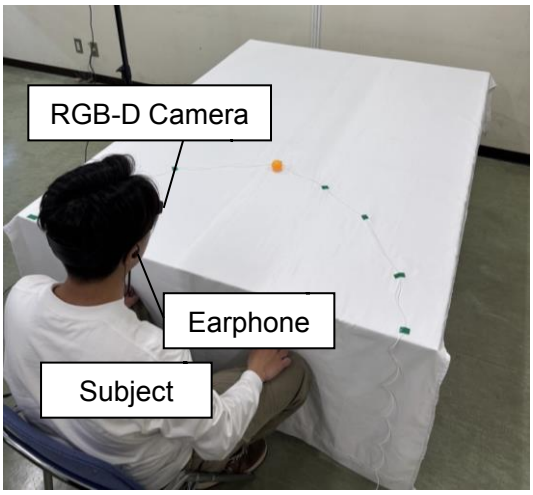


Fig. 5 Stationary Object Task (Experiment Scene)

3.1.2. Moving Object

The experimental setup for this task is shown in Fig. 6. The experiment was conducted on a table measuring 1.8 m in length and 1.2 m in width. A webcam, used to measure the distance between the ball and the racket, was mounted on a tripod 1.0 m above the table to provide a full view of the table surface. To ensure the ball rolled at a constant speed, the table was inclined at a 10-degree angle, and the ball was launched toward the subject at approximately 60 m s^{-1}. The subject's task was to predict the ball's arrival position based on audio feedback and to move a racket measuring 5 cm in width to that predicted position. The subjects were instructed to move the racket along the edge of the table.

In this task, six throwing positions were defined by dividing the experimental table equally into six sections along its length. The ball's throwing position was randomised for each trial.

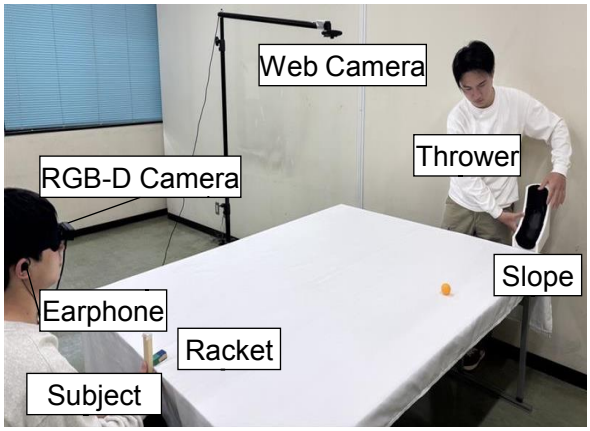


Fig. 6 Moving Object Task (Experiment Scene)

3.2. Experimental Conditions

A 2×2 experimental design was used to investigate the effects of sound type and distance model on performance and perceived usability. The different tasks were used in this study, namely, a moving object task and a stationary object task described in Section 3.1.

The two categories of sound type, continuous and intermittent, are described in Section 2.2.2. The two categories of distance model, linear and exponential, are described in Section 2.2.3. The order in which the conditions and tasks were performed was randomised.

3.3. Experimental Protocol

The subjects of this experiment were three males with an average age of 22 years. The subjects were given 5 minutes of practice time to familiarise themselves with the use of the system. For each condition, 10 trials were conducted for each task. After every condition, the subjects were then asked to respond to the SUS for subjective evaluation, and after a one-minute break,

3.4. Measurements

In this experiment, the System Usability Scale (SUS) was used as an objective distance evaluation and a subjective evaluation index to evaluate the effectiveness of the system.

3.4.1. Distance Error

The accuracy of the system was evaluated using the distance to the object in each experiment.

In the stationary object experiment, the distance between the tip of the index finger pointed by the subject and the centre of the table tennis ball was measured using a measuring tape.

In the moving object experiment, the average distance between the racket and the ball that the subject moved was used as the evaluation index. To measure the distance, the distance between the centre point of the blue part of the racket and the centre point of the ball was calculated based on the video image from the web camera installed directly above the test stand. If the racket and the ball made contact, the distance between the two points was set to 0 cm.

3.4.2. System Usability Scale (SUS)

The System Usability Scale (SUS) was used to evaluate the usability of the various conditions. SUS is a psychological scale developed by John Brook in the UK in 1996 to measure user satisfaction with a product [13]. In this experiment, the Japanese translation of SUS by Yamauchi was used [14]. The SUS questionnaire consists of 10 questions, each of which is answered on a 5-point scale, with an average score of 68.1 on a 100-point scale.

4. RESULTS

Fig. 7 presents the experimental results and SUS scores, showing the average distance error between the actual ball position and the position estimated by the user under the different conditions.

4.1. Distance Error

The results indicate that the linear distance model condition tends to produce a smaller mean error than the ex-

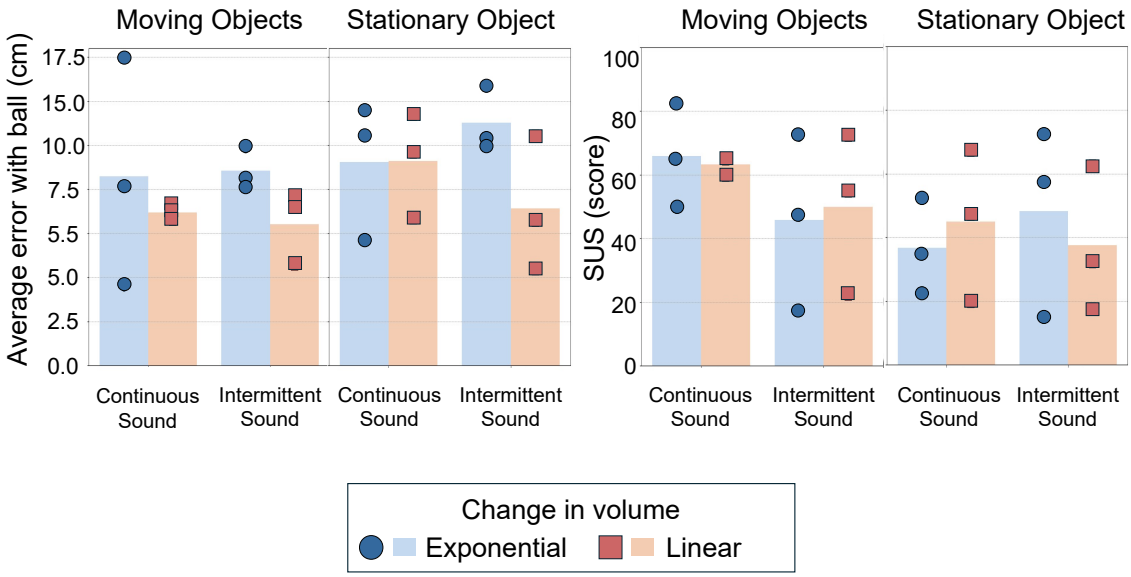


Fig. 7 Experimental Results

ponential distance model condition for both moving and stationary object tasks. Furthermore, participants showed a tendency toward lower error when using continuous sounds in the exponential distance model condition, and intermittent sounds in the linear model condition, across both tasks.

4.2. System Usability Scale (SUS)

For the moving object task, the continuous sounds appear received higher SUS scores compared to the intermittent one. However, no trend was observed regarding the distance model used (linear vs. exponential), and no trend could be observed for the effects of type of sound.

For the stationary object task, SUS scores generally aligned with the trends observed in average error. In contrast, for the moving object task, SUS scores did not show a clear correspondence with the average error.

5. DISCUSSION

This study investigated the effects of distance model (Linear vs. Exponential) and sound type (Continuous vs. Intermittent) on users' ability to localize objects using only auditory cues, without visual information. Overall, the Linear distance model tended to produce smaller mean errors compared to the Exponential model. One possible explanation is that the linear change in sound intensity may be easier for users to interpret intuitively, offering a more consistent perceptual mapping between distance and volume. In contrast, the exponential model produces gradual sound changes at longer distances and rapid changes at shorter distances, which may hinder consistent spatial recognition.

Regarding sound type, continuous tones performed slightly better in the Exponential condition, whereas intermittent tones performed slightly better in the Linear condition. This may suggest that certain sound types are more compatible with specific modulation models, potentially because users adapt more easily to certain audio dynamics. However, these trends were not strongly pronounced, and further investigation is required to determine optimal pairings of sound type and modulation method.

In terms of subjective usability (SUS scores), participants showed a slight preference for continuous sounds in the moving object task. No clear preference was observed between the Linear and Exponential models. For the stationary object task, the SUS scores generally aligned with the trends in average error, supporting the objective results. In contrast, for the moving object task, SUS scores did not correspond well with the distance error, possibly due to an increased cognitive load in dynamic scenarios and individual differences in interpreting spatial audio.

It should be noted that the number of participants in this study was limited to three, which restricts the generalisability of the findings. As such, the current results should be interpreted as indicative of preliminary trends rather than definitive conclusions. To obtain more reliable insights into the impact of acoustic conditions on auditory spatial perception, future studies must include a larger and more diverse sample and apply statistical analyses to validate observed effects.

6. CONCLUSION

This study aimed to enhance spatial perception for visually impaired individuals by proposing the use of spatial audio feedback, leveraging human sound localization abilities as a supplement to verbal descriptions. Specifically, we developed a prototype system that generates spatial audio based on the real-time positional information of a ball, tracked by a camera, relative to the user's head position.

The experiment examined the effects of two factors—sound type (continuous or intermittent) and distance attenuation model (linear or exponential)—on spatial perception for both dynamic and static objects. The results indicated that the linear attenuation model generally pro-

duced smaller average errors in position prediction for both dynamic and static objects compared to the exponential model. Additionally, a tendency was observed where continuous sound performed better in the exponential model, while intermittent sound was more effective in the linear model.

However, this study was a preliminary investigation with only three participants, and the findings should be interpreted as indicative trends rather than definitive conclusions. Future research should involve a larger and more diverse sample size, accompanied by statistical analysis, to validate the observed effects and establish a more generalizable understanding of the impact of acoustic conditions on auditory spatial perception.

REFERENCES

[1] R. Bourne et al., "Trends in prevalence of blindness and distance and near vision impairment over 30 years: An analysis for the Global Burden of Disease Study," *The Lancet Global Health*, vol. 9, no. 2, e130–e143, Feb. 1, 2021, ISSN: 2214-109X. PMID: 33275950. Accessed: May 7, 2025. [Online]. Available: https : / / www . thelancet . com / journals/langlo/article/PIIS2214-109X(20)30425-3/fulltext.

[2] *Envision App - OCR that speaks out the visual world*, https://www.letsenvision.com/app. Accessed: Apr. 16, 2025.

[3] *Envision Glasses - Smart Glasses for People who are Blind or Low Vision*, https://www.letsenvision.com/glasses/home. Accessed: Apr. 16, 2025.

[4] W. D. Marslen-Wilson, "Functional Parallelism in Spoken Word-Recognition," *Cognition*, vol. 25, no. 1–2, pp. 71–102, Mar. 1987, ISSN: 00100277. Accessed: Jan. 21, 2025.

[5] G. Fant, *Auditory Analysis and Perception of Speech*. Elsevier, Dec. 2012, ISBN: 978-0-323-14548-0.

[6] G. Yang and J. Saniie, "Sight-to-Sound Human-Machine Interface for Guiding and Navigating Visually Impaired People," *IEEE Access*, vol. 8, pp. 185 416–185 428, 2020, ISSN: 2169-3536.

[7] X. Hu, A. Song, Z. Wei, and H. Zeng, "StereoPilot: A Wearable Target Location System for Blind and Visually Impaired Using Spatial Audio Rendering," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1621–1630, 2022, ISSN: 1558-0210.

[8] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-Related Transfer Functions of Human Subjects," *Journal of the Audio Engineering Society*, vol. 43, no. 5, pp. 300–321, 1995.

[9] *Synthizer/synthizer*, synthizer, May 6, 2025. Accessed: May 7, 2025. [Online]. Available: https : //github.com/synthizer/synthizer.

[10] M. Naguib and R. Wiley, "Estimating the Distance to a Source of Sound: Mechanisms and Adaptations for Long-Range Communication," *Animal Behaviour*, vol. 62, no. 5, pp. 825–837, Nov. 2001, ISSN: 00033472.

[11] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory distance perception in humans: A review of cues, development, neuronal bases, and effects of sensory loss," *Attention, Perception, & Psychophysics*, vol. 78, no. 2, pp. 373–395, Feb. 2016, ISSN: 1943-393X.

[12] A.-p. Yang, H.-m. Hu, X. Zhang, L. Ding, and C.-K. Chen, "Natural and forced arm reach ranges in sitting position," *International Journal of Industrial Ergonomics*, vol. 86, p. 103 185, Nov. 2021, ISSN: 0169-8141.

[13] J. Brooke, "SUS - A quick and dirty usability scale,"

[14] S. Yamauchi, "Sus in welfare equipment (in japanese)," 2016.