IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 May 2024, accepted 22 June 2024, date of publication 26 June 2024, date of current version 8 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3419102

RESEARCH ARTICLE

Impact of Device and Environment on Visual-Auditory Sensory Substitution: A Comprehensive Behavioral Analysis Using the vOICe Algorithm

MOOSEOP KIM[1,2], YUNKYUNG PARK[1], KYEONGDEOK MOON[1], AND CHI YOON JEONG[1,2]

[1] Sensory Augmentation Research Section, Digital Convergence Research Laboratory, Electronics Telecommunications Research Institute (ETRI), Daejeon 34129, Republic of Korea
[2] Center of Artificial Intelligence, University of Science and Technology (UST), Daejeon 34113, Republic of Korea

Corresponding author: Chi Yoon Jeong (iamready@etri.re.kr)

This work was supported by the Electronics and Telecommunications Research Institute (ETRI) Grant funded by Korean Government (Fundamental Technology Research for Human-Centric Autonomous Intelligent Systems) under Grant 24ZB1200.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB) of Kangnam University under Application No. KNU-HR2022007.

ABSTRACT Recent studies have revealed that visual-auditory sensory substitution devices (SSDs) can effectively convey visual information to visually impaired or blind individuals through sound. However, SSDs are still not widely available to the visually impaired and blind community. Addressing these challenges requires not only the development of efficient SSD algorithms but also the evaluation of the impact of SSDs on the devices and environments in which they are used. This study represents the first attempt to analyze the impact of the device or environment used for SSDs on users' perceptual abilities. To achieve this goal, we developed an experimental procedure that involves both the training of the SSD algorithm and the changing environment and devices that receive the audio signal. Two user experiments were conducted and revealed that user perception is significantly affected by the device and environment used for SSDs. These findings underscore the importance of considering the effect of the device and environment in which it will be used when designing an SSD algorithm or training system.

INDEX TERMS Sensory substitution, visual-auditory conversion, visual perception.

I. INTRODUCTION

The lack of vision not only impacts daily life, causing mobility issues, but also limits experiences and understanding of situations. Several research studies have been conducted to convey visual information to visually impaired or blind individuals. One of the most promising methodologies is the utilization of visual-auditory sensory substitution devices (SSDs), which employ specific algorithms to convert visual data into auditory signals. Previous studies have demonstrated the potential of SSDs in supporting visually impaired individuals across various applications. For example, using visual-auditory SSDs, those with visual impairments or partial sight can effectively identify objects [1], [2], detect and

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Abdur Razzaque.

evade obstacles [3], extract depth cues and estimate distance to object [4], [5], [6], perform navigation task [7], [8], [9] and even recognize facial expressions [10], [11].

Recently, several efforts have been made to investigate the viability of employing deep learning for visual-auditory SSDs, using these methodologies to assess the quality of the substituted signal produced by the SSDs. For instance, in a study conducted by [12], deep learning was applied to assess two distinct encoding schemes for a visual-auditory SSD. In another study, Kim et al. [13] employed a cross-modal generative adversarial network (GAN) to identify the most effective auditory sensitivity, aiming to minimize transmission latency in visual-auditory sensory substitution. The same researchers further utilized deep learning techniques to optimize the sensory substitution algorithm concerning frequency range and mapping function [14].

Despite the prevalence of various technical approaches, the visually impaired and blind communities have not yet widely adopted SSDs [15], [16]. Several factors, such as their high cost, complexity of use and operation, and the time required to comprehend and adapt to the algorithms, hinder their use. Moreover, most studies have been conducted in laboratory environments, without considering the real-world conditions in which SSDs are used. Therefore, we believe that analyzing and researching the various conditions under which SSDs are used can help to close this gap.

The process of training a visual-auditory SSD entails acquiring the ability to interpret spatial patterns or environmental information conveyed through sound [4], [17], [18], [19], [20]. However, given its exclusive reliance on sound, this approach is susceptible to changes in the device delivering the audio signal and the user's auditory environment. Unfortunately, most research on visual-auditory SSDs has focused on developing effective systems without considering actual operating conditions. To assess the efficacy of a visual-auditory SSD, it is essential to conduct an evaluation examining the extent to which visual information is conveyed through the substituted audio input across diverse conditions.

Previous studies have used devices such as headphones and bone conduction headsets as visual-auditory SSDs. However, no studies have directly compared the effectiveness of these devices. Moreover, there have been no reported cases studying the effectiveness of SSDs in various living environments. Therefore, it is important to analyze changes in the extent to which users perceive visual information conveyed by substituted audio signals under various conditions, in conjunction with the development of efficient algorithms or SSDs.

To address these issues, we first developed an integrated experimental procedure to train individuals without prior experience with SSDs. We aimed to evaluate their ability to distinguish visual information from transmitted audio signals, considering changes in devices and environments. This experiment demonstrated how participants' perceptions of SSDs change under real-world use conditions. Three representative types of commercially available devices and three categories of environmental sounds in daily life were used. To the best of our knowledge, this is the first study to analyze the extent to which users visually perceive changes based on the device and environment used for SSDs. The findings of this study indicate that the development of efficient SSD algorithms requires full consideration of the conditions under which SSDs are used. Additionally, they can serve as guidelines for designing programs and platforms to train users on the effective use of SSDs. The main contributions of this study can be summarized as follows:

- We present an experimental procedure that consider both SSD algorithm training and changing device/environment conditions to verify the ability to select correct visual image corresponding to the presented substituted audio signal.
- We investigate whether the type of device used for SSD, such as earphones, headphones, or bone conduction headsets, had a significant impact on user performance.
- We show that changes in outdoor ambient sounds impact the efficiency of visual-auditory SSDs.

The remainder of this paper is organized as follows: Section II presents an overview of the related work on the training of visual-auditory sensory substitution. A detailed description of the proposed method, including the experimental setup and procedure, is given in Section III. Experimental results are presented in Section IV. The impact of device and environment and future work for the development of efficient visual–auditory SSDs are discussed in Section V. Finally, the conclusions are set out in Section VI.

II. RELATED WORK

Visual-auditory SSD enables the interpretation of visual information through auditory signals. To what extent can visual information be transmitted using SSD, and how can user efficiently learn the SSD algorithm? These issues have always been raised as major concerns because visual information is artificially coded using an audio frequency. Therefore, training visual-auditory SSD involves learning to perceive visual patterns from auditory signals.

In an early study by Amedi et al. [17], blindfolded sighted subjects were trained on pattern recognition using the vOICe (the letters in the middle of the abbreviation stand for "Oh I see") [21]. Over a period of 20 days, participants underwent training to identify a range of object positioned in front of them. At the end of the training period, participants demonstrated approximately 70% accuracy in selecting the appropriate image for a given sound. In addition to pattern recognition, Auvray et al. [18] conducted a study that demonstrated the possibility of object localization with vOICe training. They showed that blindfolded participants were able to approach a table in a room and accurately point to an object placed on the table from one of nine possible locations. In a similar study conducted by Jacomuzzi and Bruno [19], blindfolded participants were presented with one of nine rectangles at various locations on a screen.

They then listened to the corresponding sound generated by the vOICe and indicated the location in the screen where the target rectangle was located. The task resulted in a performance accuracy of around 90%. In addition, Butorova et al. [4] conducted an experiment to test the effectiveness of depth perception in visual-auditory SSD. The aim of this study was to evaluate the influence of linear perspective, one of the monocular depth cues, on the accuracy of object localization using vOICe. The study demonstrated that participants tended to overestimate distance in depth more than in width. However, the group with linear perspective exhibited greater overestimation in both depth and width. Moreover, a study by Pesnot Lerousseau et al. [20] found that participants' ability to identify sounds was influenced by visual distractors presented simultaneously, indicating shared processes between vision and sound. Additionally, participants' performance during training and their associated experiences depended on their auditory abilities.

Following the recent success of deep learning, several efforts have been proposed to enhance the efficiency of the visual-auditory SSD algorithm. The study by [13] proposed a cross-modal GAN-based evaluation method to identify the optimal auditory sensitivity for analyzing suitable auditory sensitivity in visual-auditory SSD. They modelled that the temporal length of the auditory signal for sensory substitution can be reduced by 50% using a deep learning model. The model was validated on three groups of participants (congenitally blind, late blind, and sighted users). In another study [14], a deep learning method was used to optimize the vOICe algorithm's frequency range and frequency mapping function. However, these studies only showed the possibility of applying deep learning methods to visual-auditory SSDs, and the training method was identical to existing behavioral experiments.

As described so far, behavioral findings from existing visual-auditory SSD studies have shown the potential for conveying visual information through sound. However, it is important to note that these studies have been conducted primarily in laboratory settings, without consideration of the real-world environments in which SSDs are actually used.

III. PROPOSED METHODS

A. THE VOICE ALGORITHM

We employed the vOICe algorithm, which is designed to convert and translate visual images into auditory input. As illustrated in Fig. 1, the vOICe algorithm converts visual information captured by a camera into sound using three parameters. To convert two-dimensional visual images into audio, the vOICe algorithm processes the image using column-wise scanning. The vertical position of each column is then coded into a predefined sound frequency, with a higher position corresponding to a higher frequency. The horizontal dimension of the image is mapped to time, with each column representing a moment in time from left to right, and the leftmost column of the image represents the earliest moment. Additionally, the visual intensity of each pixel in the image is converted into the corresponding volume of a sound. Thus, the conversion of a visual image into a sound occurs from left to right by adding up the sounds represented by the vertical position of all pixels in the corresponding column at a given time point.

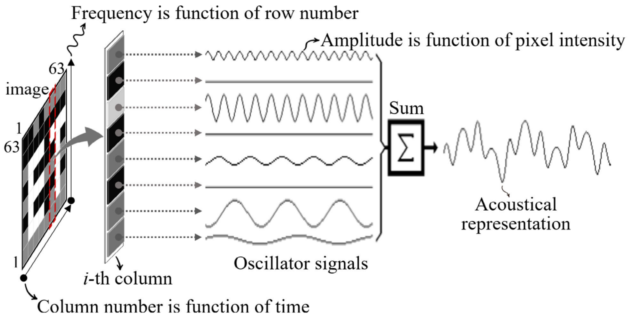
Frequency is function of row number
Amplitude is function of pixel intensity
63
image
1
63
1
Sum
Σ
Acoustical representation
Oscillator signals
i-th column
Column number is function of time

FIGURE 1. Description of vOICe conversion method. The vOICe visual-auditory sensory substitution method converts visual data into an audio signal. The Y-axis information (pixel position) of each column of the image is translated into the pitch and frequency of a sinusoidal sound. Consequently, the higher position within the column is represented by a higher pitch compared to the lower positions. The X-axis information of each column in the visual image is conveyed temporally.

B. PARTICIPANTS

A total of 45 sighted individuals with normal hearing, aged between 19 and 47 years (23 men, 22 women, mean age = 27.9 ± 7.1 years), participated in this study. All participants were unfamiliar with the vOICe algorithm and had no prior experience with any other SSDs. Although SSDs are generally aimed at supporting visually impaired and blind people, we recruited sighted participants in accordance with the recommendation in a study by [22] and [23]. All participants were randomly assigned to one of three groups based on the type of audio transmission device: wireless earphones—Sennheiser MOMENTUM True Wireless 3 Earbuds (N=15, 7 women, mean age 25.7±6.18); wireless headphones— Sony WH-1000XM5 (N=15, 8 women, mean age 30.8±7.25); and bone conduction headsets— Shokz OpenRun Pro (N=15, 7 women, mean age 27.3±7.23).

All participants provided written informed consent and received monetary compensation for their participation. The study was conducted in cooperation with the Graduate School of Welfare at Kangnam University. All research procedures and experiments were conducted in accordance with the principles of the Declaration of Helsinki. The Institutional Review Board (IRB) at Kangnam University (KNU-HR2022007) approved the study.

C. STIMULI

The visual images consisted of simple black and white shapes, including five basic shapes (circle, triangle, square, pentagon, and hexagon) and 20 variations. The auditory

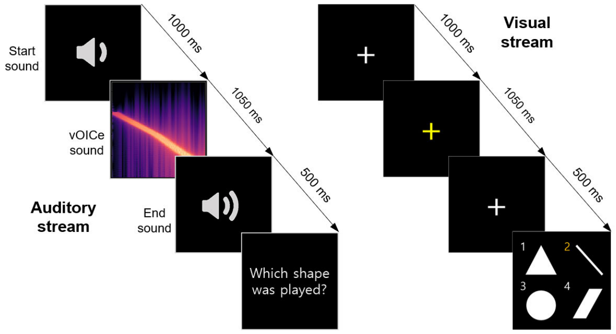

FIGURE 2. Auditory stimuli sequence for the experiment. Visual feedback was used during training to teach the participants to interpret and perceive the SSD sound of vOICe. To maintain their concentration during the experiment, visual feedback was provided using symbols and images corresponding to the onsets of the SSD sound stream. The white plus symbol in the visual stream for training and testing indicates the start and end signals and the yellow plus symbol represents the vOICe sound of the image playing.

stimuli were converted from visual images using the vOICe algorithm. A given visual image was divided into 64 rows and 64 columns of pixels. For vertical information, a column consisted of 64 pixels, each corresponding to a predefined sound frequency from a range of 80 to 7,600 Hz, in the increasing order of the Mel-scaled frequency distribution with a higher row in a higher vertical position. For horizontal information, an image divided into 64-pixel columns was scanned from left to right at the rate of 1.05 seconds per frame. The pixel sounds corresponding to 64 pixels in the same column were generated following the abovementioned vertical information conversion method and played simultaneously. The loudness of the sound was determined by the intensity of the visual image, with white producing the loudest sound and black being silent. The 16 gray levels were used to map the intermediate values between white and black.

The sequence shown in Fig. 2 was used to generate auditory stimuli for the visual-auditory conversion of a given image. To conduct the experiment in an environment similar to that of visually impaired people, all participants were blindfolded and relied on their hearing throughout the experiment. A one-second starting sound was given to participants to indicate the beginning of the visual-auditory conversion signal transmitted through the headset. After the starting sound, a converted sound for a given visual image using the vOICe followed. A short end sound followed the vOICe sound to indicate that the transmission of the visual-auditory conversion signal for the corresponding image frame had ended. Participants may have been confused with the starting point if SSD sounds were conveyed continuously. To prevent this confusion regarding the starting point, an additional 50 ms click sound was provided to inform the participants of the start of the next vOICe sound. Upon wearing the audio device, the volume was set at 15% and adjusted within ±5% upon participants' request.

D. EXPERIMENTAL SETUP

The experiment was conducted on laptop computers using off-the-shelf devices to convey visual-auditory sound. Images were transformed into artificial sounds, which were then played through the above mentioned three types of hearing devices.

We developed a program consisting of seven-step sessions to conduct the experiment and collect data from participants at each stage. Sessions were held at intervals of a minimum of two to a maximum of four days, and each participant participated in no more than one trial per day. To reduce fatigue and maintain concentration, participants were permitted to take a break at any point during the experiment. When they resumed, the experiment continued from the stopping point. However, the total time for each session did not exceed 45 minutes. Furthermore, all participant activities were automatically saved on a NoSQL server database (MongoDB) for analysis purposes.

E. EXPERIMENTAL PROCEDURE

The experimental procedure is summarized in Figure 3. Following the experimental design, the entire experiment procedure was divided into three task blocks: (1) learning the SSD algorithm, (2) inter-device discrimination task, and (3) inter-environment discrimination task. Each step of the experiment consists of a training and test stage. Before advancing to the next step, participants underwent a four-alternative forced choice test: they were presented with four images and had to select the one that corresponded to the vOICe sound they heard. This test aimed to quantitatively evaluate the extent of participants' learning during each step.

Before commencing the experimental procedure, all participants were provided with an explanation of the image-to-sound conversion rule and instructed to utilize and visualize these rules during their tasks. Subsequently, they completed the SSD algorithm learning, which constituted the first block of the task procedures, without any exposure to the vOICe sound.

The SSD algorithm learning task comprised four sessions, each with a training and test stage. The task involved three training groups, which differed with regard to the type of device used to hear auditory stimuli during training: wireless earphones (Group A), wireless headphones (Group B), and bone conduction headsets (Group C). During the training stage, we used five basic shapes and 10 out of 20 variant images (randomly selected). These 15 selected images were converted to audio 10 times in a pseudo-randomized order, resulting in a total of 150 items for training. Participants were instructed to observe changes in sound accompanying changes in the images and to imagine the relationship between the image and the related audio. Each vOICe sound was heard by all participants, followed by the display of the visual image on the screen for feedback purposes. After receiving visual feedback for the heard audio, participants continued to the next stimulus by pressing the space bar.

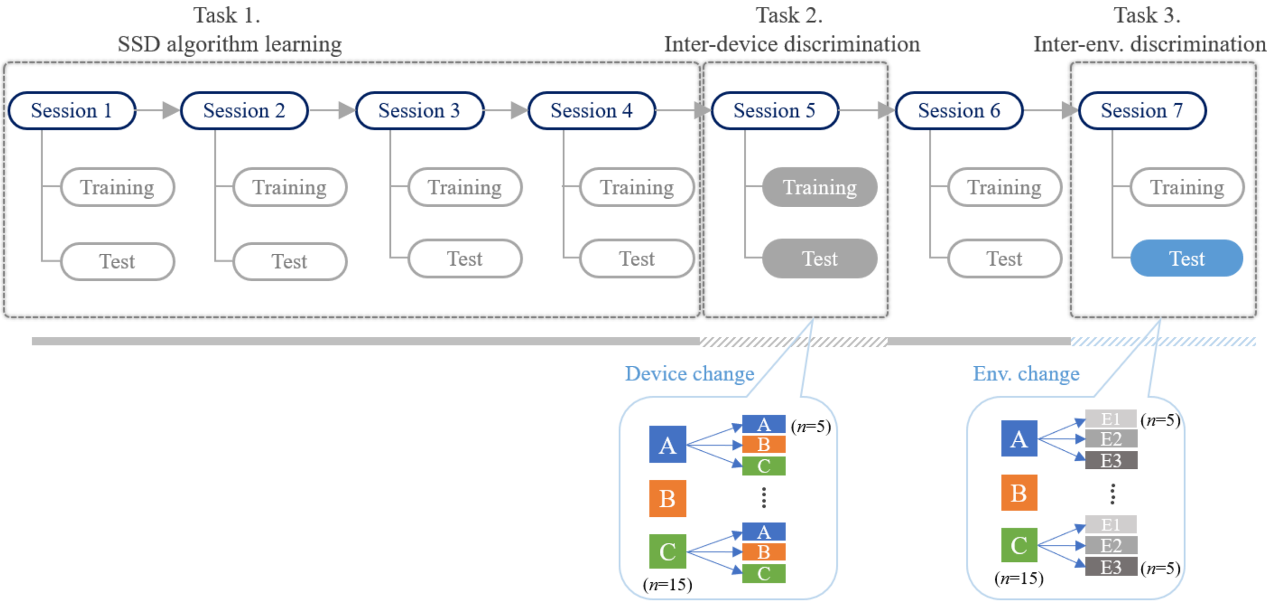

FIGURE 3. Illustration of experimental procedure. The experiment involved three groups of sighted participants, each using a different device to receive SSD sounds. All participants completed seven-step sessions, divided into three task blocks. The training stage procedure was identical for all steps, except for the randomly selected images. To examine the dependency between the device and the environmental sound, participants from each user group were randomly assigned to one of three subgroups and participated in the fifth and seventh tasks of the experiment.

After completing 150 training trials, participants were automatically directed to the test stage to evaluate their ability to identify visual images from audio stimuli. The test stage followed the same procedure as the training stage. However, all visual images were taken from the completed training stage, while the remaining images were new to the participants. During this stage, we randomly repeated the procedure five times, resulting in a total of 125 test items. For each trial of the test stage, participants were presented with an audio stimulus and four visual images and asked to select the visual image that corresponded to the heard sound. Each participant completed four forced-choice tasks, with a chance level of 0.25, to identify the correct visual image for each sound. The remaining three steps of the SSD algorithm learning task were identical, except for the shapes randomly selected for use in each training stage.

To compare the identification of SSD sounds across different experimental conditions, an inter-device identification task was conducted after the first four sessions of the procedure. Participants were instructed to train the vOICe sound corresponding to the change in the transmission device. Each participant group was randomly divided into three sub-groups according to the device used to convey sound. For example, user group A (n=15) was divided into three subgroups: AA, AB, and AC, each consisting of five participants. Consequently, the experiment was conducted with a total of nine subgroups (AA, AB, AC, BA, BB, BC, CA, CB, CC). The first letter of the group name indicates the group, and the last letter indicates the sub-group within the group using the changed device. Therefore, the AA, BB, and CC subgroups used the same device as in previous steps and served as a reference group for comparison with other subgroups within the same group. The experimental procedure for the training and test stages was similar to that of the SSD algorithm learning task.

After completing the inter-device discrimination task, participants in subgroups that used different devices were required to conduct an adaptation session where they returned to using their original device. This was completed during the sixth session of the experimental procedure and was necessary to ensure consistent results in the following steps, as variations could occur due to subjects using different devices. The experimental procedure for the training and test stages was identical to that of the previous steps.

To investigate the impact of environmental sounds on the effectiveness of SSDs, we conducted an inter-environment discrimination task. In the last session of the experimental procedure, the training stage for this step was conducted using the same procedure as in the previous steps. During the test stage, each group was divided into three subgroups were presented with three different types of environmental sounds through separate speakers (Bang & Olufsen Beosound A1). These sounds were: natural environment (E1), which included wind and rain sounds; outdoor environment (E2), which included object sounds such as of roads and crossroads; and indoor environment (E3), which included conversation and radio sounds. Throughout the experiment, environmental sounds were played at a volume ranging from

30 dB to 70 dB. The test stage of this step followed the same procedure as those of all previous steps.

IV. RESULTS

A. BEHAVIORAL PERFORMANCE

As the first part of our experimental procedure to investigate the differences in efficiency between different devices used to receive SSD sounds, we conducted SSD algorithm learning on participants with no prior experience with SSDs. This task comprised four sessions, each conducted every four days. The experiment's performance was evaluated by calculating the proportion of participants and groups that selected the correct answer during each session's testing stage.

Fig. 4 illustrates the changes in average training time and proportion of correct answers obtained during the SSD algorithm learning task. The results indicate that, despite slight differences in the change rate depending on individual characteristics, the participants' overall training time decreased as they progressed through the task. In addition, participants in both Group A, who wore wireless earphones, and Group B, who wore wireless headphones, gradually perceived visual information from the converted vOICe sounds as training progressed. No significant difference was found between the two groups. However, in the case of group C, who wore wireless bone conduction headphones, the average increased as the sessions progressed but did not reach the overall average of all user groups involved in the task.

B. EFFECT OF DEVICE CHANGE

Our study aimed to investigate the extent to which changes in the device used to receive SSD sounds affect the recognition of visual information. This task was conducted during the fifth step of the experimental procedure. To achieve this goal, we divided the participants into three subgroups, two of which were tested with different audio devices, while the third was tested without any changes to their device. The collected data were statistically analyzed using SigmaPlot 14.5 software. We evaluated how experimental participant groups perceived information differently when changing devices by calculating group-specific means and standard deviations. Statistical evaluation was conducted using one-way analysis of variance, and post-hoc verification was performed using the Bonferroni t-test. Significance was considered when the p-value was less than 0.05.

A significant difference was found in group A, in which a difference in the recognition level of subgroups was observed depending on the device used in the test, even if learning is carried out using the same device. Fig. 5 illustrates the comparison of subgroup recognition within the group, according to the change in SSDs. Among the three subgroups in Group A that used earphones, subgroup AA, which did not change the device during the test, showed a t-value of 2.98 and a significance value (p) of 0.034 compared to subgroup AC, which changed the device to bone conduction headphones. Furthermore, subgroup AB who switched to headphones, and subgroup AC who switched to bone conduction headphones exhibited a significant difference, with a t-value of 3.94 and a significance value (p) of 0.006. However, no significant difference was observed between subgroup AA and subgroup AB. Group B, who used headphones for SSD learning, also experienced a change in visual recognition when they changed devices. Subgroup BB, who did not change devices during the test, and the subgroup BC, who switched to bone conduction headphones, showed a significant difference with a t-value of 3.39 and a significance value (p) of 0.016. However, no other difference was observed in switching from headphones to other devices, other than an overall decrease in participants' perception. In Group C, where participants used wireless bone conduction headphones, the percentage of correct responses increased even after participants switched to wireless earphones (subgroup CA) and wireless headphones (subgroup CB). In all device change experiments, recognition rates increased across all experimental groups when switching to headphones. Conversely, when the device was changed to a bone conduction headset, the ability to recognize visual information decreased in all experimental groups. These results indicate that SSD performance is influenced not only by efficient algorithms but also by the specific device used.

C. EFFECT OF AMBIENT SOUNDS

Finally, to investigate a potential dependency on the environment for perceiving visual information from an audio signal, we conducted the final test stage of this experiment, presenting three types of environmental sounds to each group of participants through separate speakers. Fig. 6 depicts the change in visual perception when three environmental sounds were played to each group of participants using different devices. For Group A, who used wireless earphones for SSD learning, environmental sounds did not affect the perception of visual information, except for the indoor environment (E3). In this group, for the E1 subgroup (natural environment), the t-value was 3.08 and the significance value was 0.028, and for the E2 subgroup (outdoor environment), the t-value was 3.95 and the significance value (p) was 0.006, indicating a significant difference compared to the E3 subgroup (indoor environment). Group B, who used wireless headphones for SSD learning, showed similar results to Group A: the environmental sounds did not affect the perception of visual information, with the exception of the indoor environment (E3). In this group, only the E2 subgroup (outdoor environment) had a t-value of 2.88 and a significance value (p) of 0.042, showing a significant difference compared to the E3 subgroup. However, for Group C, who used bone conduction headsets, no significant difference was observed in any of the subgroups according to environmental sound. In particular, the other two groups whose ears were covered or blocked showed similar perceptual abilities when using

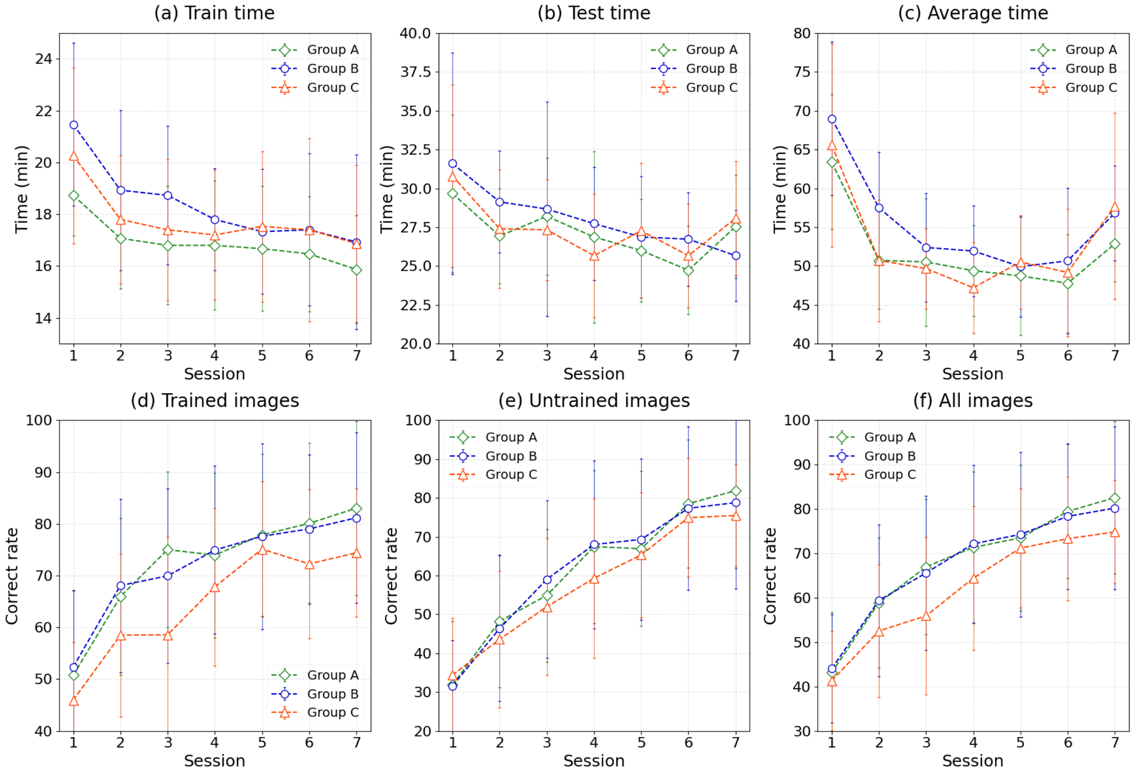

FIGURE 4. Mean performance of the three participant groups in the training sessions. Participants (*n*=15) were randomly assigned to one of three groups based on the type of audio device: wireless earphones, wireless headphones, or bone conduction headsets. The figure depicts changes in consumption time and the average proportion of correct answers as the procedure progressed. The upper row depicts the former, while the lower row depicts the latter. The results were plotted against the change in the average behavioral score across the subject groups. The following variables were measured: (a) training time of each session, (b) testing time for evaluation, (c) total execution time of each session, (d) proportion of correct answers for trained images in each session, (e) proportion of correct answers for untrained images in each session, and (f) proportion of correct answers for all images.

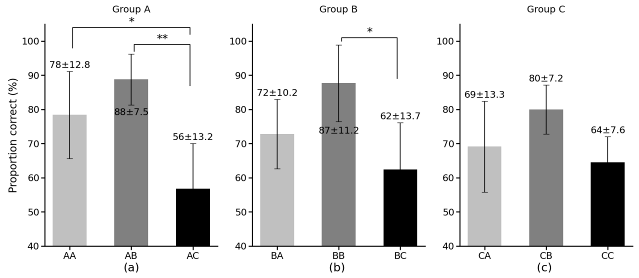

FIGURE 5. Comparison of recognition in accordance with the change of hearing devices. Each subject group was divided into three subgroups. (a) Group A trained with wireless earphones, (b) Group B trained with wireless headphones, and (c) Group C trained with bone conduction headset. The first letter of the subgroup name indicates the original group, and the last letter indicates the sub-group within the group using the changed device. Asterisks atop the bar plot indicate the significance level (* $p < 0.05$, ** $p < 0.01$).

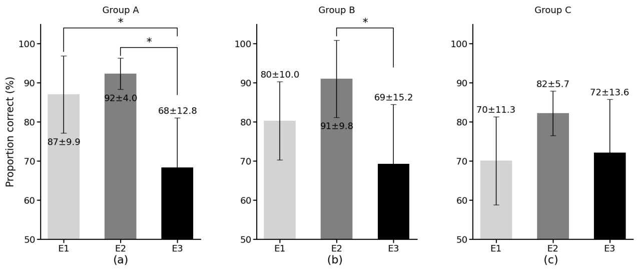

FIGURE 6. Comparison of recognition in accordance with the change of hearing environment. The experimental groups were divided into subgroups labelled E1, E2, and E3, which used sounds from natural, outdoor, and indoor environments, respectively, during the test stage. (a) Group A trained with wireless earphones, (b) Group B trained with wireless headphones, and (c) Group C trained with bone conduction headset. Asterisks atop the bar plot indicate the significance level (* $p <$ 0.05).

bone conduction headsets in E1 (natural environment) and E3 (indoor environment), which were clearly different. These results demonstrate that Group C is particularly affected by environmental sound. Our experimental results reveal differences in the visual information perceived by users depending on the environment in which the SSD is actually used.

V. DISCUSSION

A. IMPACT OF DEVICE AND ENVIRONMENT

We have demonstrated how the visual information perceived by the user from the transmitted audio signal varies depending on the SSD and environment in which it is used. To achieve our goal, we developed an experimental procedure that considers both the training of the SSD algorithm and the changing conditions of the device and environment in which the audio signal is heard.

With regard to learning the visual-auditory SSD algorithm, all participants underwent four training sessions in which they learned to identify visual image from vOICe sound. The aim of this task was to evaluate whether subjects' ability to identify visual images from audio input changed as they progressed through the learning procedure. Performance was evaluated using the trained as well as untrained images. The experimental results reveal that the learning time decreased and the user's perceptual ability increased during the learning process, regardless of the subject groups classified by the device used for training. Participants performed above chance levels on all tasks, regardless of prior knowledge of the test items. This demonstrates that visual-auditory conversion using vOICe can be effectively understood with short-time training and is highly adaptable to users with no experience of SSDs. However, some differences were observed depending on the type of device used. In particular, the group of participants using bone conduction headsets, which exposed them to external noise while listening to the audio signal, had more difficulty learning than the groups using other devices. This finding suggests that although performance improvements may occur after a short period of training, they are highly dependent on the SSD used and the ambient noise.

Assessing the impact of changing the device used to train the SSD algorithm and that used in practice on the user's perceptual abilities revealed that the impact varied depending on the device. Our experiment demonstrated that the user's overall perceptual ability increased during testing when headphones were used, regardless of the devices used in training. However, if the training was conducted using earphones or headphones but tested with a bone conduction headset, the time taken for the participant to complete the test increased and their performance significantly decreased. The reason for this result is that, although the experiment was conducted under well-controlled laboratory conditions, the vOICe sound transmitted was different from the other two devices, which may have made it difficult for participants to adapt. Another finding from this task was that the participants' performance improved when they were trained with the bone conduction headset and then tested with another device. However, this increase was not statistically significant. The experimental results reveal that learning with a bone conduction headset can provide a lower bound for recognizing information through SSD, regardless of changes in the actual device or environment. Therefore, considering the practical use of SSDs, this finding recommends using a bone conduction headset as the device for efficient SSD learning.

We were also interested in the issue of the environment of SSDs, which has not been explored before. We thus investigated the extent to which users could recognize visual information from audio as the ambient sound changed. For visual-acoustic SSDs, it is important for users to hear surrounding sounds in addition to the visual information converted to audio. Therefore, to develop efficient SSD algorithms, these situations must be considered. Thus, we examined the impact of the environment on perceptual abilities when using the SSD. The study revealed that alterations in the environment had a noteworthy effect on the learning times of the subjects. However, their perceptual abilities continued to improve. The experiment resulted in the highest and lowest performance in the outdoor and indoor environments, respectively, across all subject groups. Overall, the experiment results revealed significantly higher performance when vOICe sounds were played in conjunction with ambient sounds from natural or outdoor environments. However, the group that used bone conduction headsets did not exhibit any significant differences in the ambient sounds used during the experiment. The performance difference between subject groups seems to be attributed to the impact of external sounds, which depends on the characteristics of the device used by each group. Furthermore, superior performance in outdoor settings may be attributed to the implementation of a Mel-scale frequency mapping function in the vOICe algorithm utilized in our experiments. In addition, the study specifically revealed that alterations in the environment exerted a greater influence on learning time than did changes in the device.

B. FUTURE WORK AND LIMITATION

This study has demonstrated that the visual information perceived by the user from an audio-visual converted SSD is influenced by the device or environment in which it is used. However, there are some possible improvements that can be made to this study. First, our experiments did not include a sufficient number of participants to validate the concept. In behavioral experiments, individual participant characteristics can significantly impact the results, particularly with small sample sizes. Therefore, it is important to consider the effect of each individual's results on the overall average. This experiment's sample size of 45 participants may be small to draw conclusions. While the findings may be valid for this sample size, extending the conclusions to the general population may not be justified. Furthermore, we recruited sighted individuals as participants based on the recommendation of previous studies. Therefore, the findings of this study may not produce similar outcomes in visually impaired or blind individuals. Future work should thus expand on this study by increasing the sample size and including a more diverse group of participants for behavioral experiments.

Second, the current study used the vOICe algorithm for visual-auditory SSD. It is noteworthy that even with the vOICe algorithm used in this experiment, the user's perception may vary depending on the frequency mapping function used. While we used the Mel-scale function as the frequency mapping function for the vOICe algorithm, to obtain more definitive results, direct experiments should be performed with different frequency mapping functions, including the traditional exponential mapping function. Another possible improvement to this study would be to use different SSD algorithms for generalization, including vOICe. The main reason for choosing vOICe for this study is that it is the most widely used SSD algorithm. Existing visual-auditory SSD algorithms use different conversion methods. Because of this methodological difference, the converted sound of the visual image is qualitatively different. Therefore, this research can be extended by comparing different SSD algorithms to derive generalized results.

Finally, the current study was strictly limited by the training time. Long training times are undesirable in cognitive studies because participants may become bored and lose concentration toward the end of the session. During the four training sessions in this study, subjects were allowed to stop and restart at any time to maintain concentration, but the total pure training time for each session did not exceed 45 minutes. This study can be extended by comparing the effects of long-term training with a wide range of visual data. In addition, although research on sensory substitution has typically been conducted using behavioral experiments, which are time consuming and resource intensive, making them unfeasible for use with large numbers of participants. Recent research has proposed methods using deep learning as an alternative to these problems. Therefore, future research may also consider using deep learning method to evaluate the performance of SSD algorithms in various environments.

VI. CONCLUSION

This study investigated the effects of device and environment on the performance of visual-auditory SSDs. To achieve this objective, we extended previous research by introducing new experimental procedures and conducted behavioral evaluations to analyze the impact of changes in device and environment on user performance. Our results revealed that user performance varied significantly depending on the device used for the visual-auditory SSD and the ambient sound. Although this study is laboratory-based and limited by a small sample size, it highlights the pros and cons of the devices used in visual-auditory SSDs, and considerations for future use in real-world environments. While the use of bone conduction headsets is efficient for learning the visual-auditory SSD algorithm, our experimental results also reveal that visual-auditory SSDs are most effective in outdoor environments. However, given that behavioral experiments depend on many factors, including user characteristics and the environment, further research is clearly needed to support or refute our findings. The findings of this study, nonetheless, can guide future research in various directions, including the design of SSD algorithms and devices and the development of learning programs.

ACKNOWLEDGMENT

The authors thank the members of the Graduate School of Social Welfare, Kangnam University, for their valuable participation in the experiments.

REFERENCES

[1] G. Hamilton-Fletcher and K. C. Chan, "Auditory scene analysis principles improve image reconstruction abilities of novice vision-to-audio sensory substitution users," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 5868–5871.

[2] G. Hamilton-Fletcher, J. Alvarez, M. Obrist, and J. Ward, "SoundSight: A mobile sensory substitution device that sonifies colour, distance, and temperature," *J. Multimodal User Interfaces*, vol. 16, no. 1, pp. 107–123, Mar. 2022.

[3] D. Osinski and D. R. Hjelme, "A sensory substitution device inspired by the human visual system," in *Proc. 11th Int. Conf. Human Syst. Interact. (HSI)*, Jul. 2018, pp. 186–192.

[4] A. S. Butorova, A. A. Naizagarinova, D. A. Tarasov, and A. P. Sergeev, "The vOICe visual-auditory sensory substitution technology in the depth perception task," in *Proc. 6th Sci. School Dyn. Complex Netw. Appl. (DCNA)*, Sep. 2022, pp. 65–68.

[5] L. Commère and J. Rouat, "Evaluation of short-range depth sonifications for visual-to-auditory sensory substitution," *IEEE Trans. Human-Mach. Syst.*, vol. 53, no. 3, pp. 479–489, Jun. 2023.

[6] J. C. D. Klerk, D. Vogts, and J. L. Wesson, "Investigating techniques for gaining depth perception using visual-to-auditory sensory substitution," in *Proc. Conf. South Afr. Inst. Comput. Scientists Inf. Technologists*, Sep. 2020, pp. 141–148.

[7] A. Neugebauer, K. Rifai, M. Getzlaff, and S. Wahl, "Navigation aid for blind persons by visual-to-auditory sensory substitution: A pilot study," *PLoS ONE*, vol. 15, no. 8, Aug. 2020, Art. no. e0237344.

[8] C. Jicol, T. Lloyd-Esenkaya, M. J. Proulx, S. Lange-Smith, M. Scheller, E. O'Neill, and K. Petrini, "Efficiency of sensory substitution devices alone and in combination with self-motion for spatial navigation in sighted and visually impaired," *Frontiers Psychol.*, vol. 11, p. 1443, Jul. 2020.

[9] A. Maimon, O. Yizhar, G. Buchs, B. Heimler, and A. Amedi, "A case study in phenomenology of visual experience with retinal prosthesis versus visual-to-auditory sensory substitution," *Neuropsychologia*, vol. 173, Aug. 2022, Art. no. 108305.

[10] E. Striem-Amit, M. Guendelman, and A. Amedi, "'Visual' acuity of the congenitally blind using visual-to-auditory sensory substitution," *Seeing Perceiving*, vol. 7, no. 3, 2012, Art. no. e33136.

[11] R. Arbel, B. Heimler, and A. Amedi, "Congenitally blind adults can learn to identify face-shapes via auditory sensory substitution and successfully generalize some of the learned features," *Sci. Rep.*, vol. 12, no. 1, p. 4330, Mar. 2022.

[12] D. Hu, D. Wang, X. Li, F. Nie, and Q. Wang, "Listen to the image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7964–7973.

[13] M. Kim, Y. Park, K. Moon, and C. Y. Jeong, "Analysis and validation of cross-modal generative adversarial network for sensory substitution," *Int. J. Environ. Res. Public Health*, vol. 18, no. 12, p. 6216, Jun. 2021.

[14] M. Kim, Y. Park, K. Moon, and C. Y. Jeong, "Deep learning-based optimization of visual–auditory sensory substitution," *IEEE Access*, vol. 11, pp. 14169–14180, 2023.

[15] S. Maidenbaum, S. Abboud, and A. Amedi, "Sensory substitution: Closing the gap between basic research and widespread practical visual rehabilitation," *Neurosci. Biobehavioral Rev.*, vol. 41, pp. 3–15, Apr. 2014.

[16] L. Longin and O. Deroy, "Augmenting perception: How artificial intelligence transforms sensory substitution," *Consciousness Cognition*, vol. 99, Mar. 2022, Art. no. 103280.

[17] A. Amedi, W. M. Stern, J. A. Camprodon, F. Bermpohl, L. Merabet, S. Rotman, C. Hemond, P. Meijer, and A. Pascual-Leone, "Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex," *Nature Neurosci.*, vol. 10, no. 6, pp. 687–689, Jun. 2007.

[18] M. Auvray, S. Hanneton, and J. K. O'Regan, "Learning to perceive with a visuo-auditory substitution system: Localisation and object recognition with 'the voice,'" *Perception*, vol. 36, no. 3, pp. 416–430, 2007.

[19] A. Jacomuzzi and N. Bruno, "Perceiving occlusion through auditory–visual substitution," *Cogn. Process.*, vol. 7, no. S1, pp. 128–130, Sep. 2006.

[20] J. Pesnot Lerousseau, G. Arnold, and M. Auvray, "Training-induced plasticity enables visualizing sounds with a visual-to-auditory conversion device," *Sci. Rep.*, vol. 11, no. 1, p. 14762, Jul. 2021.

[21] P. B. L. Meijer, "An experimental system for auditory image representations," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 2, pp. 112–121, Feb. 1992.

[22] W. Kałwak, M. Reuter, M. Łukowska, B. Majchrowicz, and M. Wierzchoń, "Guidelines for quantitative and qualitative studies of sensory substitution experience," *Adapt. Behav.*, vol. 26, no. 3, pp. 111–127, Jun. 2018.

[23] E. Brulé, B. J. Tomlinson, O. Metatla, C. Jouffrais, and M. Serrano, "Review of quantitative empirical evaluations of technology for people with visual impairments," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2020, pp. 1–14.

MOOSEOP KIM received the M.S. degree in electrical engineering from Kyungpook National University, South Korea, in 1998, and the Ph.D. degree in computer science and engineering from Chungnam National University, South Korea, in 2008. He was a Research Engineer with the Organic LED (OLED) Group, Device and Materials Laboratory, LG Electronics Institute of Technology (LG Elite), Seoul, South Korea, from 1998 to 1999. He has been with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, since 1999, where he is currently a Principal Researcher. His research interests include wearable computing, activity recognition, and sensory substitution.

YUNKYUNG PARK received the M.S. degree in telecommunication engineering from Chungnam National University, South Korea, in 2006. She has been a Researcher with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, since 1987, where she is currently a Principal Researcher. Her research interests include wearable computing and sensory substitution.

KYEONGDEOK MOON received the B.S. and M.S. degrees in computer science from Hanyang University, South Korea, in 1990 and 1992, respectively, and the Ph.D. degree in information engineering from KAIST, South Korea, in 2005. From 1992 to 1996, he was a Researcher with the System Engineering Research Institute, where he worked on high-performance computing and clustering computing. Since 1997, he has been a Principal Researcher with the Electronics and Telecommunications Research Institute, where he develops home network middleware, deep learning for video analysis, and a framework for autonomously navigated ship and human augmentation architecture. His research interests include sensory substitution technology, human augmentation technology, deep learning architecture, and autonomous ships.

CHI YOON JEONG received the B.S. and M.S. degrees in electronic and electrical engineering from Pohang University of Science and Technology, Pohang, Republic of Korea, in 2002 and 2004, respectively, and the Ph.D. degree in computer science from Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea, in 2018. He is currently a Principal Researcher with the Artificial Intelligence Laboratory, Electronics and Telecommunications Research Institute, Daejeon. His research interests include computer vision, pattern recognition, machine learning, and sensory substitution.