IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 January 2023, accepted 5 February 2023, date of publication 9 February 2023, date of current version 15 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243641

RESEARCH ARTICLE

Deep Learning-Based Optimization of Visual–Auditory Sensory Substitution

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

[1]Artificial Intelligence Research Laboratory, Human Enhancement and Assistive Technology Research Section, Electronics Telecommunications Research Institute (ETRI), Daejeon 34129, South Korea
[2]Artificial Intelligence Laboratory, University of Science and Technology (UST), Daejeon 34113, South Korea

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

This work was supported by the Electronics and Telecommunications Research Institute (ETRI) Grant through the Korean Government, Fundamental Technology Research for Human-Centric Autonomous Intelligent Systems, under Grant 23ZS1200.

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 at Kangnam University under Approval No. KNU-HR2107002, and performed in line with the Declaration of Helsinki.

ABSTRACT Visual–auditory sensory substitution systems can aid visually impaired people in traveling to various places and recognizing their own environments without help from others. Although several such systems have been developed, they are either not widely used or are limited to laboratory-scale research. Among various factors that hinder the widespread use of these systems, one of the most important issues to consider is the optimization of the algorithms for sensory substitution. This study is the first attempt at exploring the possibility of using deep learning for the objective quantification of sensory substitution. To this end, we used generative adversarial networks to investigate the possibility of optimizing the vOICe algorithm, a representative visual–auditory sensory substitution method, by controlling the parameters of the method for converting an image to sound. Furthermore, we explored the effect of the parameters on the conversion scheme for the vOICe system and performed frequency-range and frequency-mapping-function experiments. The process of sensory substitution in humans was modeled to use generative models to assess the extent of visual perception from the substituted sensory signals. We verified the human-based experimental results against the modeling results. The results suggested that deep learning could be used for evaluating the efficiency of algorithms for visual–auditory sensory substitutions without labor-intensive human behavioral experiments. The introduction of deep learning for optimizing the visual–auditory conversion method is expected to facilitate studies on various aspects of sensory substitution, such as generalization and estimation of algorithm efficiency.

INDEX TERMS Cross-modal perception, generative adversarial network, sensory substitution, visual perception.

I. INTRODUCTION

Visual–auditory sensory substitution devices (SSDs) convert visual information into an auditory signal by using a specific algorithm. They can be used to provide valuable information for visually impaired people when they visit unfamiliar places or unknown environments. Previous psychophysical experiments demonstrated that SSDs could enable visually impaired and partially sighted people to recognize colors or objects [1], [2], navigate and avoid obstacles [3], [4], and even identify facial expressions [5]. These studies prove that a visual–auditory SSD is effective and practical as an alternative for providing visual information to people with impaired or defective vision; however, despite these possibilities, most visual–auditory SSDs have been restricted to laboratory-scale research and are not widely used in daily life [6].

According to previous studies [6], [7], many factors hinder the widespread use of SSDs: high cost, difficulty of use, difficulty of operation by users with visual impairment and blindness, and the long time required to adapt to these

The associate editor coordinating the review of this manuscript and approving it for publication was Felix Albu.

devices. These issues may appear trivial and distinct; however, it is necessary to closely examine all these issues comprehensively to enhance the usability of SSDs. Therefore, we decided to examine these issues by focusing on the SSD algorithm, which is a critical component because it directly affects not only the converted audio signals but also the learning result of sensory substitution. In particular, it is necessary to optimize the hyperparameters of the algorithm, as this has a major effect on the efficiency and usability of visual–auditory SSDs. vOICe (the letters in the middle stand for ''Oh I see'') [15], which is the most widely used visual–auditory SSD method, has many hyperparameters, such as the frequency range for data conversion, the type of frequency-mapping function that converts the pixel brightness into the amplitude of sound, duration, and scanning direction. However, the performance improvement is limited because the current SSDs focus on only the functionality and use a heuristic approach without performing a systematic analysis of these parameters. With human evaluation, it is difficult to systematically analyze the effects of these parameters and optimize the SSDs. Therefore, we believe that a major reason for the limited applicability of SSDs in real-world scenarios is the lack of an objective method for verification or for conducting measurements to replace human evaluation.

From the perspective of the brain, visual–auditory sensory substitution can be considered as the process of perceiving visual information through the transmitted sound. Therefore, visual perception could indicate the extent to which the participants can gain an understanding of the external world by using visual information converted into an auditory signal. Most extant studies on visual–auditory SSDs qualitatively analyzed the results of behavioral experiments or evaluated the user feedback; this is labor intensive and time consuming. Behavioral studies on visual–auditory SSDs investigated the level of task performance to estimate the progress of visual perception in specific goal-related behaviors such as object recognition [2], localization [8], and navigation [9]. However, the effectiveness of these approaches is affected by the general issue of individual sensory differences in receiving a given information and the individual effect of the environment on the experimental participants. Thus, visual perception can vary because of multiple factors such as the psychological condition of the participant, personality, time of day, weather, and other environmental elements, even under the same experimental procedures and conditions. Furthermore, it is difficult to analyze the extent to which participants can obtain visual information from behavioral studies by using a visual–auditory SSD because these SSDs differ in various aspects, including the visual data to be substituted, mapping algorithm for data conversion, and operating frequency range. Therefore, the existing methods require a large number of experimental participants to obtain the objective results of the algorithm from the subjective responses of the participants; the experiment should also be conducted under well-controlled conditions to ensure objectivity. Consequently, it is difficult to evaluate the performance of the current visual–auditory SSDs in a general and objective manner. Hence, there is an urgent need for a method that provides comparable and reliable measurements of the algorithm efficiency instead of subjective feedback to resolve these issues, although the measurements may vary owing to differences in perception between participants in some cases.

Recent advancements in deep learning have achieved remarkable success in various areas of research on artificial intelligence, thus attracting considerable attention in various domains [10]. Generative adversarial networks (GANs) [11] have gained interest as an important research topic in deep learning because they can generate synthetic samples that cannot be distinguished as true or false based on existing data. Inspired by the recent progress of GANs, several studies have focused on learning a subspace or using neural networks for the similarity measurement of different modalities. Likewise, several studies have reported the use of deep learning to automatically convert one modality information into another [12], [13], [14]. These methods focus on the possibility of conversion between modalities and do not attempt to quantify the objective measurement of visual perception or efficiency of the conversion methods of SSDs. We can expect several important advantages over previous human behavioral evaluation methods if the abovementioned characteristics of generative models are adopted for visual–auditory sensory substitution. One advantage is that we can evaluate the quality of the substituted sensory signal and thus provide information on how the signal quality can be improved. Another important advantage is the reduction in cost and effort required in experiments to evaluate the perception in visual–auditory SSDs, which can be laborious and time consuming. The final advantage is that we can objectively evaluate visual perception, which can help replace the subjective feedback from human participants during their training process.

These factors motivated us to develop an objective and systematic method for quantifying the measurement of visual perception or the efficiency of conversion methods for visual–auditory SSDs. Our hypothesis is that deep learning can provide efficient solutions for the optimization of algorithms for sensory substitution. This hypothesis can be verified by comparing the analysis results of deep learning and those of human behavioral experiments. In this study, we first modeled sensory substitution in humans using generative models to assess the ability to categorically discriminate visual information from the substituted sensory signals as an experimental challenge to provide evidence for this hypothesis. Then, we verified the results of human evaluations against the modeling results. Our method does not cover all methods of visual–auditory conversion, but allows us to test the feasibility of using deep learning for the optimization of sensory substitution algorithms with respect to frequency among various parameters. The main contributions of this study can be summarized as follows:

- We introduce a generative deep learning model that simulates the human visual system in a visual–auditory

SSD to evaluate the efficiency of converting visual information into sound.
- We investigated the visual perceptual efficiency by applying the frequency range used in vOICe to generative models; furthermore, we verified the effectiveness of the generative model through human behavioral experiments.
- We propose a novel frequency-mapping method based on the Mel-scale for the vOICe system to improve the overall quality and efficiency of the substituted auditory signal.
- We used the GAN model to analyze and compare the efficiency of the proposed Mel-scale frequency mapping with those of existing frequency-mapping methods used for vOICe and confirmed the correlation between the modeling results and user-based experimental results.

The remainder of this paper is organized as follows: Section II presents an overview of the related work on sensory substitution. Section III provides a detailed description of the proposed method. The experimental setup and evaluation results are presented in Section IV. Finally, the conclusions and future work for the development of efficient visual–auditory SSDs are discussed in Section V.

II. RELATED WORK

A. VISUAL–AUDITORY SENSORY SUBSTITUTION

Various methods are available for visual–auditory SSDs, even for converting an image into an audio stream. However, these methods differ in terms of not only the type or feature of information that is substituted but also the procedure or method used by the algorithm for converting images to sound. For example, vOICe [15] converts a visual image by using a left-to-right scan of each column, with frequency representing the vertical axis of the image and loudness representing the intensity of the pixel. Prosthesis for substitution of vision by audition (PSVA) [16] does not use column-wise horizontal scans on visual images; instead, it uses frequency increments from the bottom to the top and from the left to the right of the image and a higher density of auditory pixels at the center of the image to simulate the fovea. EyeMusic [17] allows the user to distinguish between five colors, whereas See ColOr [18] delivers seven colors. In contrast, vOICe and PSVA do not provide color information. These differences impede the generalization of visual–auditory SSDs and direct comparison of the features of each method. Therefore, a critical issue to be considered here is determining how to describe the relationship between the image-to-sound conversion method and perceptual experience reported by participants. An accurate, reliable, and unbiased method is necessary for estimating and qualifying the perceptual experience of visual information using these algorithms.

B. EVALUATION OF VISUAL PERCEPTION

The objective assessment of the perceptual experience of visual information from human behavioral observations is another aspect that needs to be seriously considered in visual–auditory sensory substitution. Many previous studies have investigated the perceptual experience of sensory substitution based on behavioral observations of the participants [19], [20], [21], [24], [25]. The participants are asked to perform tasks to extract visual information from a substituted audio signal. The performance level of the task is measured during each training session to quantitatively estimate the progress of visual perception in the behavioral experiment. The result of sensory substitution resembles the substituted modality if the participant can correctly perform a task that normally requires the substituted modality, which is vision in the case of the visual–auditory SSD. A recent study [22] argued that visual perception is not accompanied by a vision-like subjective experience even if a task that normally engages vision is accurately performed using the sensory substitution method. This implies that the quality of the perceptual experience of the visual information cannot be inferred from human behavioral experiments because a variety of qualitative differences, complexity, and each participant's distinct visual experience are not reflected in the behavioral studies. In addition, these experiments were conducted under typical experimental conditions and averaged for statistical purposes. Furthermore, the number of participants for the behavioral observation of sensory substitution in most cases was insufficient to draw a general conclusion on the perceptual experience of visual information. This scenario is more serious when the participants do not have sufficient training or practice. Therefore, human behavioral experiments on sensory substitution need to be conducted using precise and systematic methods for objectively estimating the subjective perceptual experience of visual information.

C. DEEP LEARNING FOR VISUAL–AUDITORY CONVERSION

Given the recent success of deep learning, many cross-modal learning methods have been proposed to investigate the relationship and correspondence between visual and audio modalities. An implicit method for visual–auditory sensory substitution, called the auto-encoded V2A (AEV2A) model, was recently proposed in [23]. AEV2A uses a combination of an encoder and a decoder, which consists of multiple layers of long short-term memory cells for synthesizing auditory streams from images. However, this method makes it difficult for a participant to understand the conversion scheme because there is no consistent correlation between changes in the input image and the resulting sound obtained from AEV2A. In [26], the authors employed a deep belief network with an extension-restricted Boltzmann machine to learn a shared representation between modalities; they showed that the integration of auditory and visual information into a common latent code improved speech recognition when the auditory signal was distorted. The authors of [14] proposed a model to predict sound from videos of hitting and scratching objects using a drumstick. To this end, they used a recurrent neural network to learn the sound features within a visual scene, and then produced

synthesized sounds from these features. Recently, GANs have gained considerable attention in cross-modal generation because of their exceptional performance in data synthesis. An interesting approach of generating images from audio and vice versa was first introduced in [13], which utilized this property of GANs. They used conditional GANs (CGANs) [34] to generate one modality, whereas the other modality was provided as an input. For this purpose, they used two separate image-to-sound and sound-to-image networks to perform cross-modal generation in both ways. Inspired by this study, the authors of [27] presented a unified model called cross-modal cycle GAN (CMC-GAN) to perform cross-modal mappings between image and audio. CMC-GAN comprises four encoder–decoder sub-networks (visual-to-audio, audio-to-visual, video-to-video, and audio-to-audio) to build four generation paths with four discriminators.

In [28], the authors proposed a modal translation network (MT-Net) from visual to auditory senses to generate audio descriptions from images. They demonstrated that the proposed model could generate intelligible audio descriptions from visual images to a good extent. Thus, they argued that the proposed model could assist visually impaired people in perceiving the environment better. All the abovementioned approaches express visual data through semantic audio descriptions or generate a symbolic sound that represents the image. However, it is difficult to describe new environments or the shapes of moving objects by using semantic or representative symbolic audio descriptions. According to an earlier study [29], virtual sounds are more efficient in providing guidance than semantic descriptions of speech because they can immediately convey the spatial information; this enables the efficient transfer of visual information as accurately as possible to participants.

Recently, only a few attempts have been made to utilize machine learning to estimate the auditory sensitivity and evaluate the encoding scheme of the visual–auditory SSD. These methods use deep learning to evaluate the quality of the substituted signal of the SSD; hence, they provide information on the extent to which the signal quality can be improved. For example, the authors of [12] used machine learning to examine the behavioral success using auditory encoding of a visual–auditory SSD; they proposed two cross-modal perception models for late-blind and congenitally blind individuals because their exposures to visual stimulation are different. In [30], a cross-modal GAN-based evaluation method was proposed to find the optimal auditory sensitivity to reduce the transmission latency in visual–auditory sensory substitution. However, these studies have simply demonstrated the possibility of applying deep learning methods to SSD. They did not extend the proposed approaches to a practical method to evaluate or optimize the conversion algorithm of sensory substitution.

III. PROPOSED METHOD

Research on the evaluation of visual perception or the efficiency of the conversion method for visual–auditory SSD has been conducted using human behavioral analysis. However, this method relies on the subjective feedback of participants, and, sometimes, the number of participants is insufficient to draw a general conclusion. In this section, we propose a method that uses machine learning to objectively evaluate the encoding scheme and systematically quantify the performance of visual perception of visual–auditory SSDs.

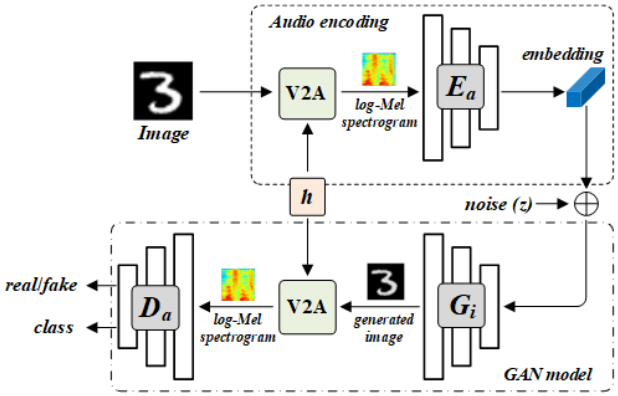


FIGURE 1. Architecture of the proposed computational model to imitate the visual perception from a substituted auditory signal. V2A represents the vOICe system adopted to convert the visual image into an audio signal. h denotes the hyperparameter applied to the vOICe system, which is related to frequency, and $\oplus$ denotes the concatenation operator. The feature embedding network (E_a) extracts an audio embedding vector from the auditory visual data. The image generator (G_i) is trained using an audio discriminator (D_a), which evaluates if the generated image is realistic and verifies that it matches the correct class.

A. ARCHITECTURE FOR OPTIMIZATION OF VISUAL–AUDITORY CONVERSION

Inspired by the outstanding ability of GAN to generate a new sample for cross-modal perception, we propose a new approach to evaluate the efficiency of the encoding scheme and qualify the visual perception of a visual–auditory SSD. Our approach is similar to that employed in [12] and [30] in that it uses vOICe as the encoding method for visual–auditory sensory substitution and creates synthesized visual data from the encoded audio signals by using a cross-modal GAN. However, we extend these approaches to obtain a method that yields an optimized design of the visual–auditory SSD algorithm and objectively quantify the visual perception from the behavioral analysis.

As shown in Fig.1, the proposed computational model comprises two parts: audio encoding, which converts visual data into audio signals and extracts features from the converted audio data, and the GAN model.

Deep learning requires sufficient data on the scenarios to be predicted; however, obtaining large amounts of image data for learning visual–auditory substitution is difficult. Manually labeling the collected samples is expensive and time consuming. Therefore, we used the MNIST handwritten digit dataset [31], which is publicly available, as the visual data for our experiment. Although the images are not of a real-world environment, they can be an appropriate alternative considering that the main objective of this study was to explore the

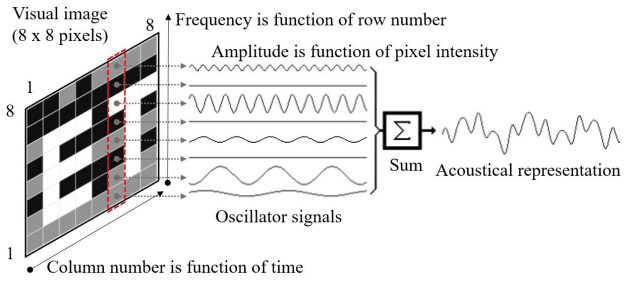


FIGURE 2. Illustration of operating principles of vOICe conversion method.

possibility of deep learning for the objective quantification of the visual perception of the substituted signal. In our experiment for evaluating the frequency hyperparameter (h), we used the MNIST dataset as the visual input to the visual-to-auditory (V2A) conversion module. We split the MNIST dataset into 60,000 training data and 10,000 testing data; then, 20% of the training data were used as the validation data. The training data were used to build the feature embedding network (E_a), the validation data were used to tune the parameters during model training, and the test data were used to evaluate the trained model.

The audio encoder comprises a visual-to-auditory (V2A) conversion module followed by an audio feature learning module that extracts the features from the encoded audio signal. We used vOICe to convert the visual images into audio signals. As shown in Fig.2, in the data conversion process of vOICe, each pixel of an image is explicitly converted into a sinusoidal wave defined by the position of the pixel and its luminance. The vertical position of a pixel in a column is coded according to a predefined frequency; the pixels have an exponential frequency distribution, with a higher pixel corresponding to a higher frequency. Therefore, the frequency range and mapping method are important factors when determining the performance of the conversion method.

In our experiment, we used the same parameters as that of the current vOICe system for performance comparison under the same conditions except for the frequency range and frequency-mapping function. The sampling frequency for the visual–audio conversion of vOICe was set to 22 kHz, and the length of the audio-sample sequence was fixed at 2 s. The visual–auditory converted signal, i.e., the output of vOICe, was transformed into a log-Mel spectrogram and then input to the feature embedding network (E_a) to extract the audio features. We used a traditional convolutional neural network (CNN) model [32] to extract the audio features; this model comprised four sequentially stacked convolutional blocks and three fully connected layers, where each convolutional block consisted of the convolutional, maxpooling, and activation layers. With the exception of the different number of filters in each convolution block, the other parameters were set to be the same.

Our approach can be used in conjunction with conditional generative models for reconstructing visual perception using features extracted from converted audio signals. Fortunately, such generative frameworks already exist in this realm, and they use additional information from both the generator and the discriminator of the GAN. The auxiliary classifier GAN (ACGAN) [35] enhances the concept of CGAN [34] by using an additional auxiliary classifier in the discriminator to classify real and generated images. Inspired by these methods, we used auxiliary classification to guide the feature embedding network (E_a) to focus on the visual perception of the converted audio signal. The image generator (G_i) uses the extracted audio features as a conditional input to generate the correct image from the converted audio signal. The image generator of the GAN model creates fake image samples from a 100-dimensional random input noise vector with a uniform distribution and 128-dimensional audio features. These features are provided as inputs to the dense layer of the generator, which contains 6,272 neurons. The output of this layer is reshaped to a two-dimensional data feature of size (7×7, 128). This reshaped feature goes through two up-sampling modules, followed by a 5×5 convolution that halves the number of feature channels, BatchNorm layers, and ReLU activation, where each layer upsamples the feature maps by double at every step. Therefore, the output-image size of the generator becomes 28×28. The generated image, which is the output of the image generator, is converted into an audio signal using vOICe and represented as a 2d log-Mel spectrogram before being input to the audio discriminator (D_a). The audio discriminator consists of four convolution layers with a 3×3 kernel and 64, 128, 256, and 512 filters, each followed by a maxpooling layer. The hidden layers of the discriminator are activated by LeakyReLU. Then, two fully connected layers are used to estimate the probability distribution over both sources and categories. Thus, the discriminator has two output branches: a binary classification with sigmoid activation to predict real or fake data, and a Softmax activation to predict multiclass classification.

B. FREQUENCY RANGE OPTIMIZATION FOR VISUAL–AUDITORY SSD

One of the main challenges in developing a visual–auditory SSD is the mismatch in the amount and characteristics of information that can be transferred through the visual and audio channels. The conversion scheme of vOICe is based on the long-evidenced correspondence between vision and audio, for example, the mapping of brightness in an image to the loudness of sound. According to previous research [36], the audible frequency range of human hearing is 20 Hz–20 kHz; however, this range is not used for frequency mapping of vOICe. For visual–auditory conversion, vOICe uses the frequency range of 500–5000 Hz, and the frequency mapping of 64 vertical pixels of the image is assigned through exponential scaling.

The evaluation of the effectiveness of different frequency ranges used for visual–auditory conversion requires an extensive user-based experiment. However, it is difficult to guarantee that the changed range of mapping frequencies is the most

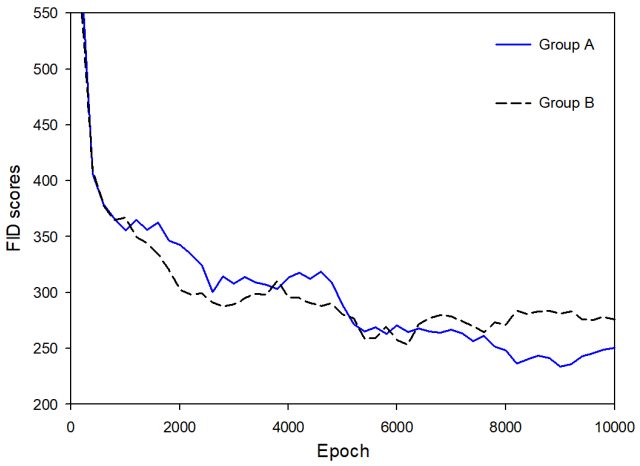


FIGURE 3. Performance of the generative model in terms of the Fréchet inception distance (FID) with respect to the training iteration for different frequency ranges of Groups A (80–7600 Hz) and B (500–5000 Hz). Each data point indicates the averaged value of 20 repetitions of the FID score to obtain reasonably good statistics in each frequency range.

effective. Therefore, it is necessary to determine an efficient approach to evaluate the efficiency of visual–auditory SSD without expensive and time-consuming human behavioral experiments. We explored whether computational models can estimate the effect of changes in the frequency range used for visual–auditory sensory substitution. We used two different frequency ranges employed in the conventional vOICe to evaluate how visual perception varies according to the change in the frequency range of vOICe. One was the frequency range of 500–5000 Hz used by the current version of vOICe. The other was the frequency range of 80–7600 Hz, which was used in earlier versions of vOICe. These two frequency ranges were used as values of the hyperparameter (h) for the visual–auditory conversion module in the audio-encoding part of the proposed framework, as indicated in Fig. 1.

We evaluated the generated images using two frequency ranges (500–5000 Hz and 80–7600 Hz) of the visual–auditory conversion in the vOICe system to achieve this purpose. Visual examination is the simplest and most direct method to observe changes in the generated image. However, the changes in the quality of the generated images should be systematically analyzed, as it is difficult to evaluate subtle changes in the generated image. We adopted the Fréchet inception distance (FID) to evaluate the quality of the generated image instead of the inception score used in [12]. This was because the FID assesses the hidden activations from the end of the classifier instead of assessing the quality of the generated samples whereas the inception score evaluates the distribution of the class probabilities of the generated images. Therefore, a lower FID score indicates that the GAN model can generate a better image from the converted audio. The program for all experiments was implemented using Python version 3.5, and the generative model was implemented using the Keras framework based on TensorFlow. All experiments were conducted on an RTX 3090 GPU card. We measured the change in the FID score for each of the 200 iterations to estimate the quality of the image generated using the aforementioned two frequency ranges.

Fig. 3 illustrates the performance trends in terms of the training epochs and FID scores of the images generated using the proposed model on the MNIST dataset for different frequency ranges. The results in Fig. 3 indicate that it is more efficient to use a wide frequency range when converting visual information into an audio signal because the FID score of Group A (80–7600 Hz) was significantly lower than that of Group B (500–5000 Hz). This indicates that the visual–auditory conversion method using a wide frequency range converted the visual information into an audio signal more efficiently than when using a narrow frequency range.

C. FREQUENCY-MAPPING OPTIMIZATION FOR VISUAL–AUDITORY SSD

The method for mapping the visual information to sound is another important factor to consider when designing a visual–auditory SSD. This is because vOICe maps the vertical position of the visual image into a range of audio frequencies. vOICe represents a direct and simple method to filter out important visual information by using an exponential distribution of the frequency mapping and to convert visual information into auditory information. It maps the output frequency such that higher pixel positions in the scanning column of the input image have higher frequencies.

$$f_e(i) = 2\pi \cdot F_L \cdot (1.0 \cdot F_H/F_L)^{(1.0 \cdot i/(M-1))}, \quad (1)$$

where M, F_L, F_H, and i represent the pixel number of a column or the height of the image, lower boundary of the frequency range, upper boundary of the frequency range, and pixel position in a column of the image, respectively.

A recent study [12] drew attention to the experimental results [33] that humans are most sensitive to sound frequencies between 2 kHz and 5 kHz. They suggested a method for adjusting the quality of the audio signal in the sensitive frequencies of the human ear. They argued that the center of an image is more important than its other regions. Based on this hypothesis, they proposed a rectified Tanh distribution for the frequency mapping of vOICe.

$$f_t(i) = 2\pi \cdot F_L + 2\pi(r/2 \cdot tanh(\alpha \cdot (i - M/2)) + r/2), \quad (2)$$

where r denotes the frequency range of $F_H - -F_L$, and α denotes the scaling parameter set to 0.06. They showed that in the evaluation of their conversion method, both the results of the inception-score computation [39] and human-based experiments were equally improved when compared with those of the existing vOICe algorithm. However, it is unclear whether this improvement in visual perception was due to the actual effect of the conversion method because their experiment was conducted with only three human participants. Furthermore, it is unclear whether the inception score alone is sufficient to evaluate the performance of the generator. Although the inception score is commonly employed in

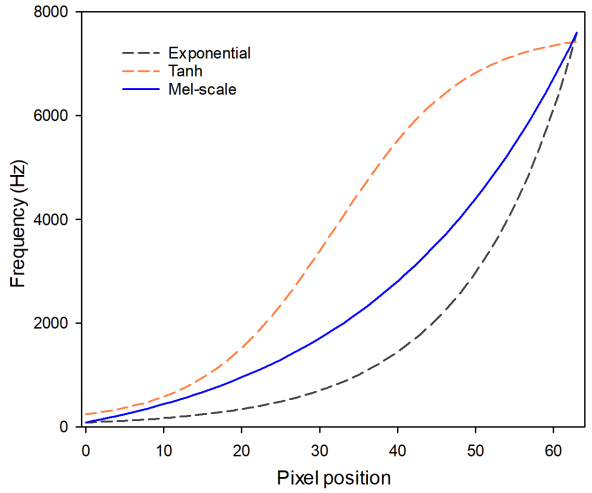


FIGURE 4. Frequency-mapping functions for mapping the visual–auditory conversion applied to the vOICe system. The frequency-mapping functions depicted are Exponential [15], Tanh [12], and the proposed Mel-scale.

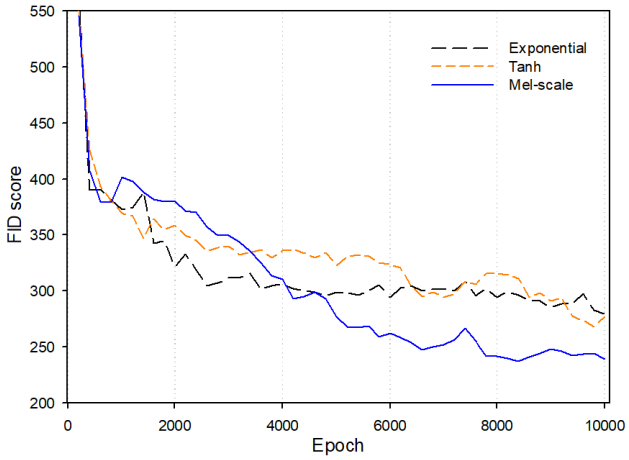


FIGURE 5. Performance of visual perception using the generative model in terms of the FID for various frequency-mapping functions: Exponential [15], Tanh [12], and the proposed Mel-scale. Each data point indicates the mean value of the FID score repeated ten times to obtain reasonably good statistics.

studies on image generation, it does not penalize a model that memorizes the training data or fails to generalize.

Recently, Mel-spectrograms [37] have been widely used in machine learning tasks such as speech recognition, acoustic scene classification, and audio-based emotion recognition. In comparison with raw waveform audio, it has the capability to reduce the amount of data that must be processed by the neural network; furthermore, it scales the actual frequencies according to the human hearing system. Motivated by these advantages of the Mel spectrogram, we propose a new frequency-mapping scheme to enhance the performance of the vOICe system. To this end, we modified the frequency range used in the vOICe system to the Mel-scale frequency banks obtained by applying the widely used Mel-frequency transformation [38].

$$\begin{aligned} M_H &= 2595 \cdot log_{10}(1 + F_H/700) \\ M_L &= 2595 \cdot log_{10}(1 + F_L/700) \\ f_m(i) &= 2\pi \cdot 700 \cdot (10^{(M_L + i(M_H - M_L)/M)/2595} - 1), \end{aligned} \quad (3)$$

where M_H and M_L represent the center frequencies of the upper and lower filter banks of the Mel-scale, corresponding to the highest and lowest frequencies (F_H, F_L) of the frequency range applied to the vOICe system, respectively. Fig. 4 shows the Mel-scaled frequency-mapping function and the frequency-mapping functions used in the vOICe system to map the vertical position of an image to the frequency of the auditory signal. As indicated in Fig. 4, the three types of frequency-mapping functions show different characteristics based on the pixel position of the image.

The visual perception of participants in the vOICe system is expected to be affected by the algorithms or by converting translating schemes substituted to substituting modality, as expected from our experiment on the frequency range. We examined the effectiveness of the visual–auditory conversion method by applying deep learning to sensory substitution to examine whether the change in the mapping frequency for visual–auditory conversion of vOICe causes a difference in visual perception.

We verified the generated images according to the change in the frequency distribution for the visual–auditory conversion scheme of the vOICe system prior to human evaluation. We used the same generative model as explained in the experiment on the frequency range; the only difference was that we converted the visual data into auditory signals by using the three different frequency-mapping functions depicted as h in Fig. 1 on the frequency ranges obtained from our experiments.

Fig. 5 shows the change in the FID score of the generated MNIST digits based on the variation in the mapping frequency for the conversion scheme of the vOICe system. It can be seen from the figure that the results of the mapping frequencies indicate similar changes in performance. However, a remarkable difference was observed in the visual recognition of the image generated by the generator as the training progressed. According to the modeling results, the frequency-mapping function using the Mel-scale proposed in this study provides better visual perception than the previous method of Exponential [15] and Tanh [12] in sensory substitution.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The participants' feedback or task performance is used as an indicator to assess the efficiency of the conversion method in the conventional assessment [17], [40]. We designed two experiments to examine the relationship between visual perception measured using the deep learning model and task performance of human behavioral observation. The first experiment was a preliminary study that compared the results of a deep learning model with the results of a human behavior experiment to find a more effective frequency range

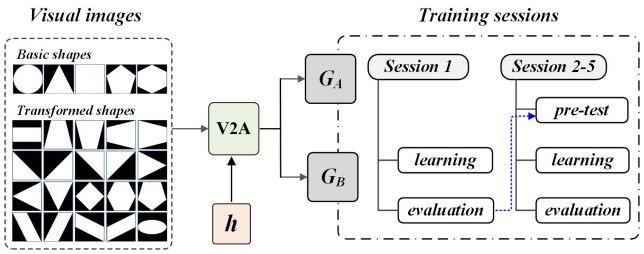


FIGURE 6. Illustration of the experimental procedure for human evaluation. *h* denotes the hyperparameters of frequency for the vOICe system. Group A (G_A) and B (G_B) refer to the group of participants who were trained with the converted sound with different frequency ranges: Group A (80–7600 Hz) and Group B (500–5000 Hz).

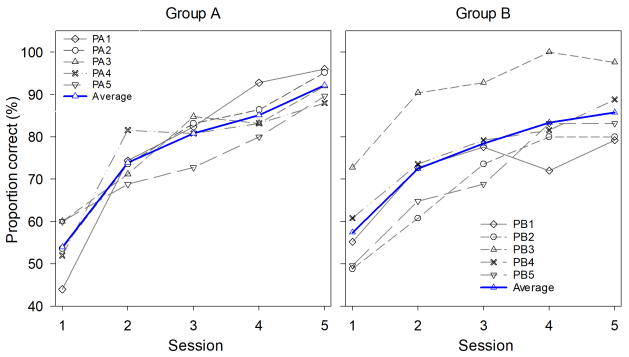


FIGURE 7. Mean performance on tested items with respect to the training session, measured in terms of proportion of correct responses. Each data point on the blue lines indicates the average score of the subject group (*n*=5) during the evaluation phase of each training session.

for visual–auditory conversion in the vOICe system. The second experiment involved a comparison of the extent of visual perception, both in the deep learning model and in the participant-based experiment, between conventional frequency distributions and frequency mapping using the Mel-scale in the vOICe system.

A. FREQUENCY-RANGE OPTIMIZATION

We hypothesized that a change in the frequency range for converting visual data affects the perceptual experience of visual information because vOICe matches the pixel position of the incoming image with the frequency of the audio output. To test this hypothesis, we converted the visual image into sound using two types of frequency ranges in the vOICe system. We expected better performance over a wide range of frequencies because it can provide greater discriminatory capacity. Therefore, the user can efficiently distinguish the auditory signal.

To this end, we conducted preliminary experiments on human participants to observe differences in the visual information obtained from the vOICe sounds converted with different frequency ranges. In [22], the authors recommended recruiting naive-sighted participants to minimize the difference based on the visual experience of participants and for ensuring smooth progress of the experiment. In the pilot study, to verify the experimental results of human evaluations against the modeling results, ten sighted subjects with normal hearing, aged between 20 and 27 years (male: female = 5: 5, mean age = 22.3 ± 2.16 years), were recruited; the subjects were tested in cooperation with the Graduate School of Welfare, Kangnam University. All the experiments in this study were conducted according to the principles of the Declaration of Helsinki and approved by the Institutional Review Board at Kangnam University (KNU-HR2107002). All participants were unaware of the purpose of the study, and none of them had prior experience with an SSD. All the participants provided written informed consent and received monetary compensation for their participation.

Fig. 6 shows the procedure of the human behavior experiment to compare the effectiveness of the frequency ranges used for visual–auditory conversion. The participants were randomly assigned to two experimental groups to train with different audio streams converted using different frequency ranges: Group A (G_A, 80–7600 Hz) and Group B (G_B, 500–5000 Hz). We used 25 simple white-colored shapes on a black background (5 basic shapes and 20 variants of these shapes) as visual stimuli to achieve the purpose of the experiment. Auditory stimuli were generated by vOICe by using different frequencies on a log-linear scale to translate the visual input into auditory output. The participants listened to the generated auditory signals using Sennheiser IE300 earphones. Fig. 6 shows that the experiment comprised five sessions conducted at intervals of a minimum of two to a maximum of four days. All participants participated in only one trial per day. Except for the first session, which consisted of the learning and evaluation phases, each training session comprised the pre-test, learning, and evaluation phases. In the pre-test phase, we programmed a total of 25 images, which were repeated randomly, twice for 50 test items, to evaluate the extent to which the participants preserved the visual experience trained in the previous session. The visual images used in the learning phase of each session comprised the five basic shapes and ten randomly selected images of the transformed shapes. During the learning phase, the selected 15 images were repeatedly converted to audio for training in a pseudo-randomized order 10 times for a total of 150 training items per session. The participants were instructed to imagine the relationship between the image and the related audio by observing changes in sound accompanying the changes in the images. The evaluation phase followed the same procedure as the pre-test phase; however, all images were used, which included the images used for training in the learning phase and images not used for training, yielding a total of 125 test items per session. Each participant group tested their perceptual experience by identifying the correct visual images based on their corresponding sounds through four forced-choice tasks, with a chance level of 0.25.

The behavioral performance was measured by calculating the proportion of correct answers for each test item. Fig. 7 shows the mean proportions of correct answers obtained by the participant groups for different frequency ranges in each training session. The perception of visual information

through the audio signal gradually increased as the training progressed. Our experimental results indicated that the participants in Group A, who trained using a wide range of frequencies, showed a steady increase in visual perception as the training progressed. This phenomenon is different from that of the participants in Group B, in which the same learning process was performed, but the improvement in the learning rate increased moderately from that in session 3. Therefore, our results show that converting visual information using the frequency range of 80–7600 Hz in SSDs is more efficient than that using 500–5500 Hz for perceiving visual information from audio signals. From our preliminary experiment, we confirmed that the analysis results with deep learning were similar to the results of the user-based experiments.

B. FREQUENCY-MAPPING OPTIMIZATION

According to the modeling results of the three frequency-mapping functions, we expect that the proposed frequency mapping using Mel-scale distributions can provide better visual perception than the existing frequency mapping methods. However, the question remains as to whether model-based analysis has the same results in the behavioral experiment conducted on human participants as in our preliminary experiments.

To examine this question, we recruited 30 volunteers aged 19–30 years (male: female = 16: 14, mean age = 23.83 ± 2.83 years) for the human evaluation experiment. All the participants provided written consent at the beginning of the first session, and none of them had participated in the preliminary experiment. The participants were randomly assigned to one of three groups according to the frequency-mapping functions for visual–auditory conversion: Group A (exponential), Group B (Tanh), and Group C (Mel-scale).

The training procedure was similar to that of the preliminary experiment illustrated in Fig. 6, except that the pre-test phase was not conducted and an additional session was added to the learning procedure. The same sets of training and test items used in the preliminary experiment were applied in each session. The learning phase of each session involved the presentation of visual images and their converted sound. In the evaluation phase of each session, four forced-choice trials were conducted; the task in each trial was to select the correct visual image corresponding to the sound heard by the participants from among four visual images.

The performance of the behavioral experiment was measured by calculating the proportion of correct choices made by the participants and groups in the evaluation phase of each training session. Fig. 8 shows the change in the average training time and proportion of correct answers for each training session. The upper part of Fig. 8 presents a plot of the change in the average time taken by the participant in each group in the learning phase as the training session progressed. Although there were slight differences in the change rate depending on the individual characteristics, the overall training time of the participants decreased as the training session progressed, regardless of the conversion scheme. However, the extent to which the participants in each group perceived the visual information from the converted vOICe sounds increased gradually, as shown in the lower portion of Fig. 8; the pattern and extent of change differed between the groups. We found that each group of participants perceived the visual image from the given auditory sound, regardless of whether the converted image was used for training. However, as the training progressed, the extent of perceived visual information from the audio heard by each participant group differed according to the frequency-mapping function. Our experimental results showed that after a given training session, the extent of perceived visual information of the group trained through Mel-scaled frequency mapping was the greatest, with an average increase rate of 10.2%, whereas the Exponential [15] and Tanh [12] mapping showed average increase rates of 7.56% and 8.54%, respectively.

When considering the average increase rate, Tanh frequency mapping showed better performance than exponential frequency mapping, but this group of participants had the lowest overall performance in visual information perception. The behavioral results in Fig. 8 show that the proposed Mel-scaled frequency mapping achieved the best performance when compared with the existing conversion methods in terms of perceiving the visual information. This finding suggests that the analysis results using the deep learning model and results obtained from human behavioral analysis are similar, as observed in the preliminary experiment.

Our experimental results demonstrated that the proposed method could reliably estimate the efficiency of visual perception under changes in the visual–auditory conversion scheme in the vOICe system without relying on subjective human evaluation. Thus, deep learning-based analysis can be an acceptable alternative for improving the objectiveness in the field of sensory substitution, which has traditionally relied on subjective human feedback for performance analysis.

C. DISCUSSION

We demonstrated that deep learning can be used to evaluate the perception of converted sensory information instead of relying on human participants and to optimize the sensory substitution algorithm by optimizing different parameters.

Although our results are promising, further studies are required to validate our findings. First, it was difficult to obtain a sufficient number of participants within a strict experimental period because of the COVID-19 pandemic. Furthermore, we did not conduct experiments to test whether the proposed method has an acceptable level of usability for other SSD systems because our study aimed to verify the possibility of using deep learning in the field of sensory substitution. However, it was revealed that our results of applying the frequency-range and frequency-mapping function for visual–auditory conversion, obtained from the machine learning model, to the vOICe system were substantially consistent. Therefore, the findings of our experiments can be extended to various attempts for generalizing the application of deep learning to the field of sensory substitution in the future.

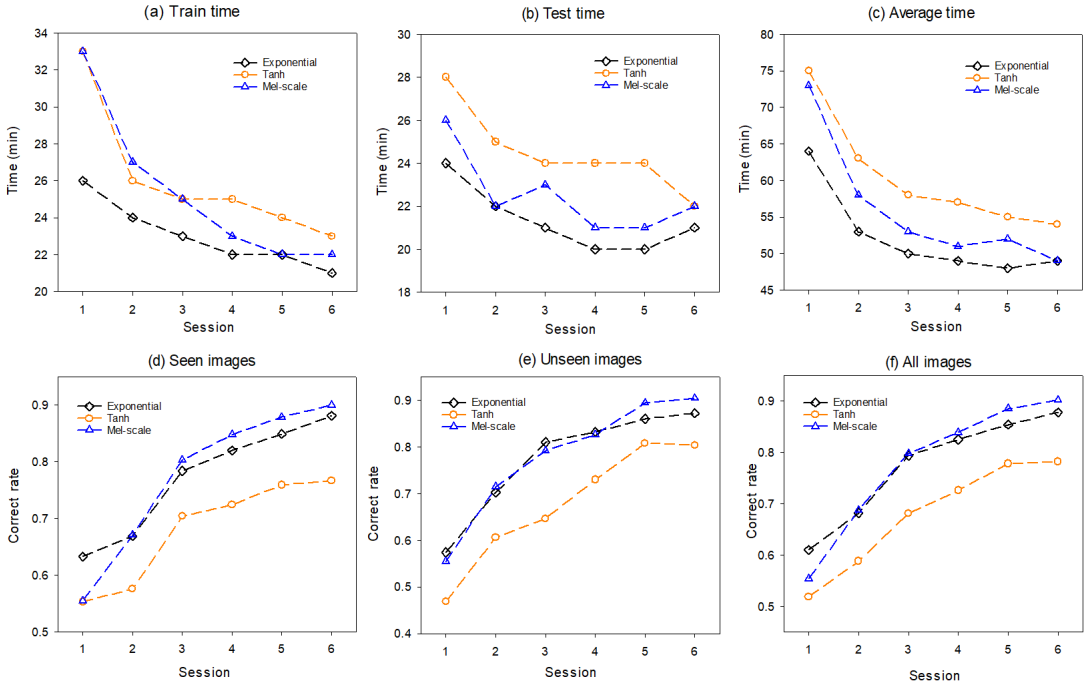


FIGURE 8. Mean performance of three participant groups in the training sessions. Each participant group (n=10) was randomly assigned to receive audio signals converted by one out of three mapping functions: Exponential [15], Tanh [12], and the proposed Mel-scale. The upper row of the figure represents changes in the training time and the lower row represents the average proportion of correct answers as the training progressed. Each result was plotted against the change in the average behavioral score across the subject groups. (a) Training time of each session; (b) testing time for evaluation; (c) total execution time of each session; (d) proportion of correct answers for images used for training in each session; (e) proportion of correct answers for images not used for training in each session; (f) proportion of correct answers for all images.

Our experiment was conducted with participants of a specific age group. However, the listening capability and sensitivity may differ depending on the age. Therefore, it is necessary to expand the experimental group to include a more diverse age range.

SSD devices aim to effectively convey visual information so that users can adapt well to their new environment. For the generalization of visual–auditory SSD, we need to validate our approach by extending it to various SSDs with different conversion schemes because we only used the vOICe system in our experiments. Our approach provides an opportunity to examine whether deep learning model-based analysis of visual perception can be used as an efficient alternative for the analysis of human behavioral experiments. However, in visual–auditory substitution, various methods that can code visual information [15], [16], [17], [18], [23], [24], [25] are available. These methods have different characteristics such as converting visual information into a substituted signal, transmitting converted signals using a target sensory modal, and presenting auditory stimuli-converted visual information. To the best of our knowledge, no previous study has directly compared these methods. Therefore, our study can be extended and applied as a general solution to verify the effectiveness of these sensory substitution methods having different characteristics.

V. CONCLUSION

A new method to objectively evaluate the effectiveness of SSD was developed in this study. We showed that introducing deep learning to the field of sensory substitution could enable the objective evaluation of the efficient frequency range and frequency-mapping functions for audio conversion without labor-intensive human behavioral experiments. To the best of our knowledge, this is the first attempt at evaluating the effectiveness of the visual—auditory conversion method via the application of a deep learning model in sensory substitution and at examining the correlation between model-based analysis and the results of human behavioral experiments.

In a preliminary experiment, we explored the possibility of using deep learning to evaluate the quality of the converted signal in sensory substitution. The results of the deep learning analysis conducted in our preliminary experiment showed that the use of a wide range of frequencies for the conversion scheme of vOICe is more efficient for obtaining visual perception from sound. Furthermore, the results of the model-based analysis were similar to those of the analysis of the human behavior experiments.

We expanded the results of the preliminary experiments to confirm whether deep learning-based analysis can be effectively applied to improve the efficiency of the visual–auditory conversion scheme in SSDs. We conducted an experiment

wherein the Mel-scale distribution function was applied in addition to the two types of frequency-mapping functions used in the existing vOICe conversion method. The experimental results showed that the Mel-scale distribution function proposed in this study had the best performance for visual perception in both model-based analysis and human behavior experiments.

Although we did not conduct a thorough experiment and verification for a large number of parameters, we demonstrated that introducing deep learning to the field of sensory substitution could enable the objective evaluation of the efficient frequency range and estimation of frequency mapping for audio conversion; this can result in the improvement of the perceptual experience. Thus, we believe that machine learning can present a technological and conceptual shift that can solve the limitations of conventional methods in sensory substitution. Although our study demonstrated the potential possibility of model-based evaluation for sensory substitution, further research is required to enhance and generalize the results of our findings. Because we used a few basic shapes as visual sensory information in the human behavior experiments in a well-controlled laboratory environment, our approach cannot be extended to training in more natural environments and is therefore not feasible for a holistic SSD. Therefore, this issue can be an interesting direction for future research. Another important research direction would be to perform an objective performance analysis of various existing techniques. Existing visual–auditory substitution methods mainly differ in terms of the conversion method. Because of this difference, a direct comparison of the existing methods is not possible and has never been attempted. Moreover, the perceptual differences due to the differences in the conversion method are unknown. Therefore, our approach of using a deep learning model for sensory substitution provides a groundwork for future research.

ACKNOWLEDGMENT

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

REFERENCES

[1] D. Osiński, M. Łukowska, D. R. Hjelme, and M. Wierzchoń, "Colorophone 2.0: A wearable color sonification device generating live stereosoundscapes—Design, implementation, and usability audit," *Sensors*, vol. 21, no. 21, p. 7351, Nov. 2021.

[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 Interface*, 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. Hum. Syst. Interact. (HSI)*, Jul. 2018, pp. 186–192.

[4] 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.

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

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

[7] A. Kristjánsson, A. Moldoveanu, O. I. Jóhannesson, O. Balan, S. Spagnol, V. V. Valgeirsdóttir, and R. Unnthorsson, "Designing sensory-substitution devices: Principles, pitfalls and potential," *Restor. Neurol. Neurosci.* vol. 34, no. 5, pp. 769–787, 2016.

[8] S. Levy-Tzedek, S. Hanassy, S. Abboud, S. Maidenbaum, and A. Amedi, "Fast, accurate reaching movements with a visual-to-auditory sensory substitution device," *Restorative Neurol. Neurosci.*, vol. 30, no. 4, pp. 313–323, 2012.

[9] C. Stoll, R. Palluel-Germain, V. Fristot, D. Pellerin, D. Alleysson, and C. Graff, "Navigating from a depth image converted into sound," *Appl. Bionics Biomech.*, vol. 2015, pp. 1–9, Jan. 2015.

[10] S. Chung, C. Y. Jeong, J. M. Lim, J. Lim, K. J. Noh, G. Kim, and H. Jeong, "Real-world multimodal lifelog dataset for human behavior study," *ETRI J.*, vol. 44, no. 3, pp. 426–437, Jun. 2022.

[11] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 23, 2022, doi: 10.1109/TKDE.2021.3130191.

[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. 7972–7981.

[13] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proc. Thematic Workshops ACM Multimedia*, Oct. 2017, pp. 349–357.

[14] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2405–2413.

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

[16] C. Capelle, C. Trullemans, P. Arno, and C. Veraart, "A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 1, pp. 1279–1293, Oct. 1998.

[17] S. Abboud, S. Hanassy, S. Levy-Tzedek, S. Maidenbaum, and A. Amedi, "EyeMusic: Introducing a 'visual" colorful experience for the blind using auditory sensory substitution," *Restor. Neurol. Neurosci.*, vol. 32, pp. 247–257, Jan. 2014.

[18] G. Bologna, B. Deville, T. Pun, and M. Vinckenbosch, "Transforming 3D coloured pixels into musical instrument notes for vision substitution applications," *EURASIP J. Image Video Process.*, vol. 2007, pp. 1–14, Dec. 2007.

[19] 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.

[20] 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. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–14.

[21] 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, pp. 1–11, Jul. 2021.

[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] V. Tóth and L. Parkkonen, "Autoencoding sensory substitution," 2019, *arXiv:1907.06286*.

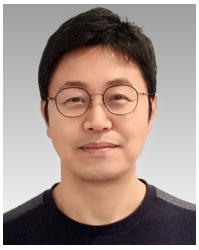
[24] S. Tian, M. Zheng, W. Zou, X. Li, and L. Zhang, "Dynamic crosswalk scene understanding for the visually impaired," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1478–1486, 2021.

[25] R. K. Katzschmann, B. Araki, and D. Rus, "Safe local navigation for visually impaired users with a time-of-flight and haptic feedback device," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 583–593, Mar. 2018.

[26] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, Bellevue, WA, USA, Jul. 2011, pp. 689–696.

[27] W. Hao, Z. Zhang, and H. Guan, "CMCGAN: A uniform framework for cross-modal visual-audio mutual generation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Feb. 2018, 6886–6893.

[28] H. Ning, X. Zheng, Y. Yuan, and X. Lu, "Audio description from image by modal translation network," *Neurocomputing*, vol. 423, pp. 124–134, Jan. 2021.
[29] J. M. Loomis, J. R. Marston, R. G. Golledge, and R. L. Klatzky, "Personal guidance system for people with visual impairment: A comparison of spatial displays for route guidance," *J. Vis. Impairment Blindness*, vol. 99, no. 4, pp. 219–232, Apr. 2005.
[30] 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.
[31] A. Baldominos, Y. Saez, and P. Isasi, "A survey of handwritten character recognition with MNIST and EMNIST," *Appl. Sci.*, vol. 9, no. 15, p. 3169, Aug. 2019.
[32] M. Kim and C. Y. Jeong, "Label-preserving data augmentation for mobile sensor data," *Multidimensional Syst. Signal Process.*, vol. 32, no. 1, pp. 115–129, Jan. 2021.
[33] S. A. Gelfand, *Essentials Audiology*, 4th ed. New York, NY, USA: Thieme Medical Publishers Inc, 2016.
[34] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[35] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Sydney, NSW, Australia, Aug. 2017, pp. 2642–2651.
[36] X. Zhao, X. Wang, and D. Cheng, "A model of co-saliency based audio attention," *Multimedia Tools Appl.*, vol. 79, nos. 31–32, pp. 23045–23069, Aug. 2020.
[37] N. Y.-H. Wang, H.-L.-S. Wang, T.-W. Wang, S.-W. Fu, X. Lu, H.-M. Wang, and Y. Tsao, "Improving the intelligibility of speech for simulated electric and acoustic stimulation using fully convolutional neural networks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 184–195, 2021.
[38] J. Olivares-Mercado, G. Sanchez-Perez, G. Aguilar-Torres, H. Perez-Meana, K. Toscano-Medina, and M. Nakano-Miyatake, *Multidimensional Features Extraction Methods in Frequency Domain*. London, U.K.: INTECH Open Access Publisher, 2012.
[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," 2016, *arXiv:1606.03498*.
[40] N. R. B. Stiles and S. Shimojo, "Auditory sensory substitution is intuitive and automatic with texture stimuli," *Sci. Rep.*, vol. 5, no. 1, pp. 1–14, Oct. 2015.



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 at the Device and Materials Laboratory, Organic LED (OLED) Group, 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 current 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 working as a Researcher with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, since 1987, where she is currently a Principal Researcher. Her current 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 at 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 the home network middleware, deep learning for video analysis, framework for autonomously navigated ships, 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 the Pohang University of Science and Technology, Pohang, South Korea, in 2002 and 2004, respectively, and the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2018. He is currently a Principal Researcher with the Artificial Intelligence Laboratory, Electronics and Telecommunications Research Institute, Daejeon. His main research interests include computer vision, pattern recognition, machine learning, and sensory substitution.