Listen to the Image

Di Hu, Dong Wang, Xuelong Li, Feiping Nie,* Qi Wang
School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University
hdui831@mail.nwpu.edu.cn, nwpuwangdong@gmail.com,
xuelong_li@ieee.org, feipingnie@gmail.com, crabwq@gmail.com

Abstract

Visual-to-auditory sensory substitution devices can assist the blind in sensing the visual environment by translating the visual information into a sound pattern. To improve the translation quality, the task performances of the blind are usually employed to evaluate different encoding schemes. In contrast to the toilsome human-based assessment, we argue that machine model can be also developed for evaluation, and more efficient. To this end, we firstly propose two distinct cross-modal perception model w.r.t. the late-blind and congenitally-blind cases, which aim to generate concrete visual contents based on the translated sound. To validate the functionality of proposed models, two novel optimization strategies w.r.t. the primary encoding scheme are presented. Further, we conduct sets of human-based experiments to evaluate and compare them with the conducted machine-based assessments in the cross-modal generation task. Their highly consistent results w.r.t. different encoding schemes indicate that using machine model to accelerate optimization evaluation and reduce experimental cost is feasible to some extent, which could dramatically promote the upgrading of encoding scheme then help the blind to improve their visual perception ability.

1. Introduction

There are millions of blind people all over the world, how to help them to "re-see" the outside world is a significant but challenging task. In general, the main causes of blindness come from various eye diseases [35]. That is to say, the visual cortex should be largely unwounded. Hence, it becomes possible to use other organs (e.g., ears) as the sensor to "visually" perceive the environment according to the theory of cross-modal plasticity [4]. In the past decades, there have been several projects attempting to help the disabled to recover their lost senses via other sensory channels,
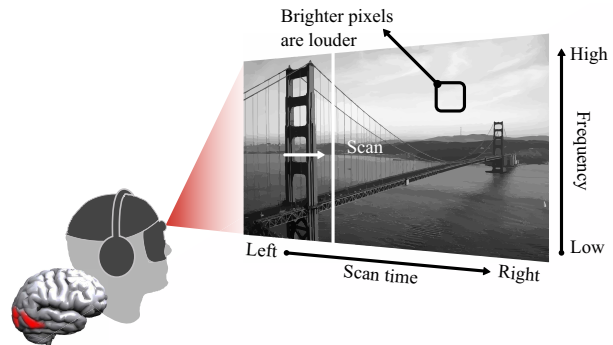

Figure 1. An illustration of the vOICe device. A camera mounted on forehead captures the scene in front of the blind, which is then converted to sound and transmitted to headphones. After 10-15 hours of training, the regions of visual cortex become active due to cross-modal plasticity. Best viewed in color.

and the relevant equipments are usually named as *Sensory Substitution* (SS) devices. In this paper, we mainly focus on the visual-to-auditory SS device of vOICe[1] (The upper case of OIC means "Oh! I See!").

The vOICe is an auditory sensory substitution device that encodes 2D gray image into 1D audio signal. Concretely, it translates vertical position to frequency, left-right position to scan time, and brightness to sound loudness, as shown in Fig. 1. After training with the device, both blindfolded sighted and blind participants can identify different objects just via the encoded audio signal [33]. More surprisingly, neural imaging studies (e.g., fMRI and PET) show that the visual cortex of trained participants are activated when listening to the vOICe audio [1, 28]. Some participants, especially the late-bind, report that they can "see" the shown objects, even the colorful appearance [18]. As a consequence, the vOICe device is considered to provide a novel and effective way to assist the blind in "visually" sensing the world via hearing [36].

However, the current vOICe encoding scheme is identi-

---

[1]www.seeingwithsound.com

fied as a primary solution [33]. More optimization efforts should be considered and evaluated to improve the translation quality. Currently, we have to resort to the cognitive assessment of participants. However, two characteristics make it difficult to be adopted in practice. First, the human-based assessment consists of complicated training and testing procedures, during which the participants have to keep high-concentration and -pressure for a long period of time. Such toilsome evaluation is unfriendly to the participants, and it is also inefficient. Second, amounts of control experiments are required to provide convincing assessments, which therefore need much efforts as well as huge employment cost. As a result, there is few works focusing on the optimization of current encoding scheme, while it indeed plays a crucial role in improving the visual perception ability of the blinds [33, 18].

By contrast, with the rapid development of machine learning techniques, the artificial perception models have taken the advantages of efficiency, economical, and convenience, which exactly settle the above problems of human-based evaluation. Hence, can we design proper machine model for cross-modal perception like human, and view it as an assessment reference for different audio encoding schemes? In this paper, to answer these questions, we develop two cross-modal perception models w.r.t. the distinct late- and congenitally-blind case, respectively. In the late-blind case, as they have seen the environments before they were blind. The learned visual experiences could help them imagine the corresponding objects when listening to the translated sound via the vOICe [1, 18]. Based on this, we propose one *Late-Blind Model* (LBM), where the visual generative model is initialized from abundant visual scene for imitating the conditions before blindness, then employed for visual generation based on the audio embeddings of separated visual perception. While the congenitally-blind have never seen the world before, therefore the absence of visual knowledge makes them difficult to imagine the shown objects. In this case, we propose a novel *Congenitally-Blind Model* (CBM), where the visual generation depends on the feedback of audio discriminator via a derivable sound translator. Without any prior knowledge in the visual domain, the CBM can generate concrete visual contents related to the audio signal. Finally, we employ the proposed model to evaluate different encoding schemes presented by us and compare the assessments with sets of conducted human-based evaluations. Briefly, our contributions are summarized as follows,

- We propose a novel computer-assisted evaluation task about the encoding schemes of visual-to-auditory SS devices for the blind community, which aims to vastly improve the efficiency of evaluation.
- Two novel cross-modal perception models are developed for the late-blind and congenitally-blind condi-

tions, respectively. The generated proper visual content w.r.t. the audio signal confirm their effectiveness.
- We present two novel optimization strategies w.r.t. the current encoding schemes, and validate them on the proposed model.
- We conduct amounts of human-based evaluation experiments for the presented encoding schemes. The highly consistent results with the machine-model verify the feasibility of machine-based assessments.

2. Related Works

2.1. Sensory Substitution

To remedy the disabled visual perception ability of the blind, many attempts have been made to covert the visual information into other sensory perception, where the visual-to-auditory sensory substitution is the most attractive and promising fashion. The early echolocation approaches provide a considerable conversion manner [16, 17], but the required sound emission makes it impractical in many scenarios, such as the noisy environment. Another important attempt is to convert the visual information into speech, e.g., the Seeing-AI project[2]. However, such direct semantic description is just developed for the low vision community, as it is hard to imagine the concrete visual shape for the blind people, especially the congenitally blind ones. And it also suffers from limited object and scene categories of the training set. Hence, more researchers suggest to encode the visual appearance into sequential audio signal according to specific rules, such as the vOICe.

Most researchers consider that cross-modal plasticity is the essential reason to the success of the SS devices [1, 21, 5, 34, 29], which is also the neurophysiological criterion for judging the effectiveness of these devices [28]. Cross-modal plasticity refers that the sensory deprivation of one modality can have an impact on the cortex development of remaining modalities [4]. For example, the auditory cortex of deaf subjects is activated by visual stimulate [9], while the visual cortex is activated by auditory messages for the blind people [34]. To explore the practical effects, Arno *et al.* [2] proposed to detect the neural activations via *Positron Emission Tomography* (PET) when the participants are required to recognize patterns and shapes with the visual-to-auditory SS devices. For both blindfolded sighted and congenitally blind participants, they found a variety of cortex regions related to visual and spatial imagery were activated relative to the baselines of audition. More related works can be found in [1, 21]. These evidences show that regions of the brain normally dedicated to vision can be used to process visual information passed by sound [36, 25]. In addition, the above studies also show that there exit effective interactions among modalities in the brain [4, 14].

---

[2]https://www.microsoft.com/en-us/seeing-ai

In view that cross-modal plasticity provides a convincing explanation for sensory substitution, how to effectively utilize it for the disabled people in practice becomes the key problem. The visual-to-auditory SS device of vOICe has provided one feasible encoding scheme that has been verified in the practical usage. But it is difficult to guarantee that whether the current scheme is the most effective one. By conducting sets of control experiments to the current scheme, such as reversing the encoding direction (e.g., the highest pitch is set to the bottom instead of the top of image), Stiles *et al.* [33] found that the primary scheme did not hold the best performance in the task of matching images and sounds, which indicates that better encoding scheme is expected to improve the cross-modal perception ability of the blind people further. However, the evaluation of encoding schemes based on human feedback is relatively complicated and inefficient. In this paper, a kind of machine model is proposed to accelerate the assessment progress, which is also convenient and economical.

2.2. Cross-modal Machine Learning

In the machine learning community, it is also expected to build effective cross-modal learning model similar to the brain. And many attempts have been made in different cross-modal tasks. Ngiam *et al.* [24] introduced a novel learning setting to effectively evaluate the shared representation across modalities, named as "hearing to see". Thereinto, the model was trained with one provided modality but tested on the other modality, which later confirmed the existing semantic correlation across modalities. Inspired by this, Srivastava *et al.* [32] proposed to generate the missing text for a given image, where the shared semantic provided the probability to predict corresponding descriptions. Further, Owens *et al.* [27] focused on the more complicated sound generation task. They presented an algorithm to synthesize sound for silent videos about hitting objects. To make the generated sound realistic enough, they resorted to a feature-based exemplar retrieval strategy instead of direct generation [27]. While more recent work proposed an immediate way in generating natural sound for wild videos [38]. On the contrary, Chung *et al.* [7] proposed to generate visually talking face based on an initial frame and speech signal. For a given audio sequence, the proposed model could generate a determined image frame that best represents the speech sample at each time step [7]. Recently, the impressive *Generation Adversarial Networks* (GANs) [12] provide more possibilities in the cross-modal generation. The early works focused on the text-based image generation [30, 37], which were all developed based on the conditional GAN [22]. Moreover, it becomes possible to generate images that did not exist before. Chen *et al.* [6] extended above model into more complicated audiovisual perception task. The proposed model [6] showed noticeable

performance in generating musical sounds based on the input image and vice versa.

Although the cross-modal machine learning models above have shown impressive generation ability for the missing modality, they do not satisfy the blind condition. This is because the blind people cannot receive the visual information, while the above models have to utilize these information during training. By contrast, we focus on this intractable cross-modal generation problem, where the missing modalities are unavailable in the whole lifecycle.

3. Cross-modal Machine Perception

3.1. Late-blind Model

People who had visual experience of outside world but go blind because of diseases or physical injury are called late-blind people. Hence, when wearing the SS device of the vOICe to perceive objects, the pre-existing visual experience in their brain can provide effective reference for cross-modal perception. The relevant cognitive experiments also confirm this, especially some blind participants could unconsciously color the objects while the color information was not encoded into sound [18]. This probably comes from the participants' memory of object color [18]. Such significant character inspires us to build one three-stage *Late-Blind Model* (LBM), as shown in Fig. 2. Concretely, we propose to model the late-blind case by decoupling the cross-modal perception into separated sound perception and out-of-sample visual perception, then coupling them for visual generation. In the first stage, as the category labels of translated sound are available, the audio convolutional network of VGGish [13] is employed as the perception model to learn effective audio embeddings via a typical sound classification task, where the input sound are represented in log-mel spectrogram[3]. For the visual modality, to achieve diversiform visual experience, the perception model is trained to model abundant visual imageries via adversarial mechanism in the second stage, which aims to imitate the circumstances of before going blind. Further, the learned audio embeddings are viewed as the input to visual generation, and the whole cross-modal perception model is fine-tuned by identifying the generated images with an off-the-shelf visual classifier[4]. Obviously, the generated images w.r.t. the translated sound should be much similar to the ones used to train the visual generator, which accordingly provides the possibilities to automatically color the shown objects.

Results. To evaluate the proposed LBM, we start with the simple handwritten digits generation task. Concretely, the MNIST images [20] are firstly translated into sounds via the vOICe for training the audio subnet and the cross-modal

---

[3] The mel-scale depicts the characteristics of human hearing.
[4] The visual classifier is trained with the same dataset used in the second stage but fixed when fine-tuning.
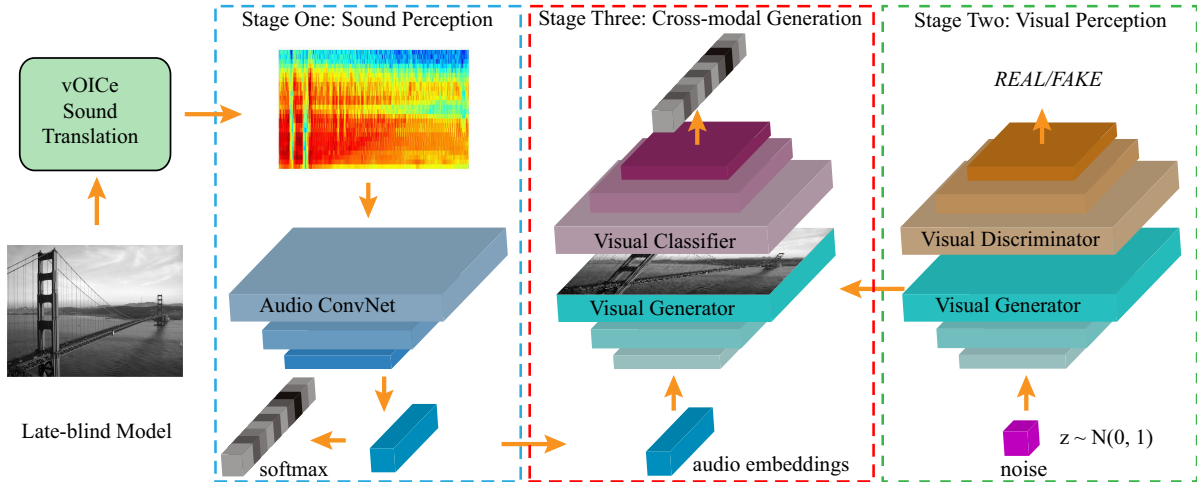
Figure 2. The diagram of the proposed late-blind model. The image and vOICe translator outside the dashed boxes indicate the circumstance of blindness when using SS devices, while the three-stage perception model within the boxes is constituted by sound embedding, visual knowledge learning, and cross-modal generation. Best viewed in color.

generation model, while the more challenging dataset of *Extension MNIST* (EMNIST) Digits [8] is employed for training the visual generator and classifier. And the official training/testing splits on both datasets are adopted [20, 8]. Fig. 3 shows some generated digit examples w.r.t. the translated sounds in the testing set of MNIST[5]. Obviously, the generated digit contents can be easily recognized and also corresponding to the labels of translated sounds, which confirm-

---

[5]Network settings and training details are in the materials.
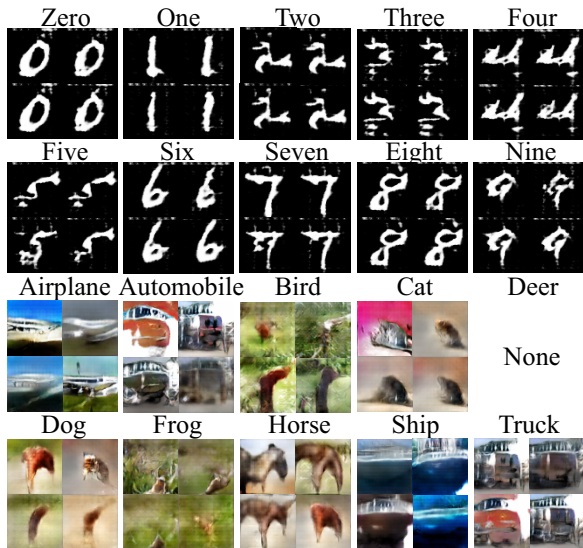

Figure 3. Generated visual examples using our late-blind model. Top two rows are digit images generated from the translated sound of MNIST and visual knowledge of EMNIST, while bottom two rows are object images generated from the sound of CIFAR-10 and visual knowledge of ImageNet.

s that the visual experience learned from EMNIST indeed helps to build visual content with audio embeddings. Further, we attempt to generate realistic objects by training LBM with more difficult datasets, i.e., using CIFAR-10 [19] for cross-modal generation and ImageNet [10] for visual knowledge learning. To improve the complexity of visual experiences, apart from the nine categories of CIFAR-10[6], we randomly select another ten classes from ImageNet for training the visual networks. As expected, although the translated sounds do not encode the color information of original object, the generated objects are automatically colored due to the priori visual knowledge from ImageNet, which provides a kind of confirmation to the theory of experience-driven multi-sensory associations [18]. But on the other hand, due to the absence of real images, our generated images are not as good as the ones generated by directly discriminating images. It exactly confirms the difficulty of cross-modal generation in the blind case.

3.2. Congenitally-blind Model

Different from late-blind people, congenitally-blind people were born blind. Their absent visual experience makes them extremely difficult to imagine the visual appearance of shown objects. However, cross-modal plasticity provides the possibility to effectively sense concrete visual content via the audio channel [18, 36], which depends on specific image-to-sound translation rules. In practice, before training the blinds to "see" objects with the vOICe device, they should learn the translation rules firstly by identifying simple shapes, which could make them sense the objects more precisely [33]. In other words, the cross-modal translation

---

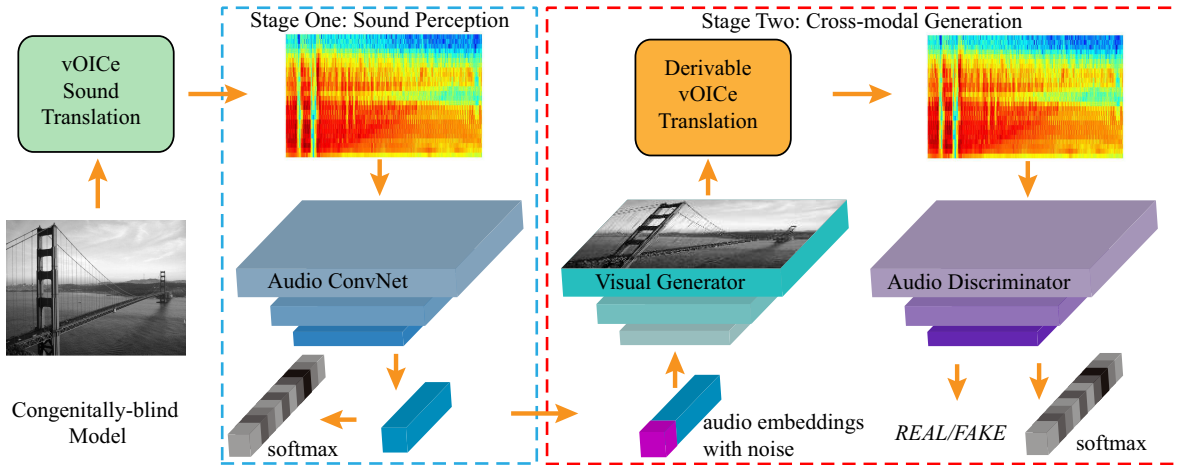[6]The deer class of CIFAR-10 is removed due to its absence in the ImageNet dataset.

Figure 4. The diagram of the proposed congenitally-blind model. The image and vOICe translator outside the dashed boxes represent the circumstance of blindness with SS devices, while the two-stage model within the boxes consists of preliminary sound embedding and cross-modal generative adversarial perception.

helps to bridge the visual and audio perception in the brain. Based on this, we propose a two-stage *Congenitally-Blind Model* (CBM), as shown in Fig. 4. Similarly with LBM, the VGGish network is firstly utilized to model the translated sound via a classification task, and the extracted embeddings are then used as the conditional input to cross-modal generation. In the second stage, without resorting to the prior visual knowledge, we propose to directly generate concrete visual contents by a novel cross-modal GAN, where the generator and discriminator deal with different modalities. By encoding the generated images into the sound modality with derivable cross-modal translation, it becomes feasible to directly compare the generated visual image and the original translated sound. Meanwhile, to generate correct visual content to the translated sound, the visual generator takes the audio embeddings as the conditional input and the audio discriminator takes the softmax regression as an auxiliary classifier, which accordingly constitute a variant auxiliary classifier GAN [26].

Results. As the visual experience is not required in the congenitally-blind case, the datasets employed for pre-training the visual generator are not needed any more. Hence, our proposed CBM is directly evaluated on MNIST and CIFAR-10 by following the traditional training/testing splits in [20, 19]. As the vOICe translation just deals with gray images, the images discriminated by sound do not contain RGB information, as shown in Fig. 5. Obviously, these image samples on both datasets are not as good as the ones of LBM, due to the compressed visual content by the sound translator and the absent visual experience. Even so, our CBM can still generate concrete visual content according to the input sound, e.g, the distinct digit forms. As the objects in CIFAR-10 suffer from complex background that distracts the sound translation for objects, the generated appearances
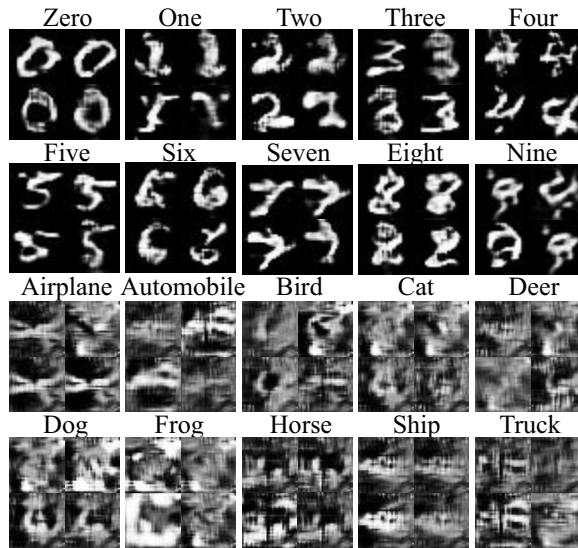

Figure 5. Generated visual examples using our congenitally-blind model. Top two rows are digit images generated from the translated sound of the MNIST dataset, while bottom two rows are object images generated from the sound of the CIFAR-10 dataset.

become low-quality while the outlines can be still captured, such as horse, airplane, etc. On the contrary, clean background can dramatically help to generate high-quality object images, and more examples can be found in the following sections.

4. Evaluation of Encoding Schemes

In view that Stiles *et al.* has shown the current encoding scheme can be further optimized to improve its applicability for the blind community [33], it becomes indispensable to efficiently evaluate different schemes according to the task

performance of cross-modal perception. Traditionally, the evaluation has to be based on the inefficient participants' feedback. In contrast, as the proposed cross-modal perception model has shown impressive visual generation ability, especially the congenitally-blind one, it is worth considering whether the machine model can be used to evaluate the encoding scheme. More importantly, such machine-based assessment is more convenient and efficient compared with the manual fashion. Hence, in this section, we make a comparison between machine- and human-based assessment, which is performed with modified encoding schemes.

Different from the simple reversal of the encoding direction [33], we aim to explore more possibilities in optimizing the primary scheme. First of all, a well-known fact is that there exist large differences in bandwidth between vision and audition [18]. When 2D images are projected into 1D audio sequence with limited length, amounts of image content and details are compressed or declined. One direct approach is to increase the audio length, which accordingly makes the vOICe encode more detailed visual contents with more audio frames. Meanwhile, the blind people can also have more time to imagine the corresponding image content. But on the other hand, such augmentation cannot be unlimited for efficiency. Hence, the time length is doubled from the primary setting of 1.05s to 2s in the first modified encoding scheme.

Apart from the bandwidth discrepancy, another crucial but previously neglected fact should be also paid attention to. Generally, humans are most sensitive (i.e. able to discern at the lowest intensity) to the sound frequencies between 2K and 5K Hz [11]. As the blind participants perceive the image content via their ears, it is necessary to provide high-quality audio signal that should exist in the sensitive frequency area. However, due to the bandwidth discrepancy between modalities, it is difficult to precisely perceive and imagine all the visual contents in front of the blind via the translated sound. To address this issue, we argue that the center of the visual field should be more important than other visual areas for the convenience of practical use. Hence, we aim to effectively project the central areas to the sensitive frequencies of human ears. Inspired by [15], a novel rectified tanh distribution is proposed for *Position-Frequency* (PF) projection, i.e.,

$$frequency = s/2 \cdot \tanh\left(\alpha \cdot (i - rows/2)\right) + s/2, \quad (1)$$

where s is the frequency range of the encoded sound, α is the scaling parameter, i is the position of translated pixel, and $rows$ is the image height. As shown in Fig. 6, it is obvious that the image centers (row 20-40) fall into the most sensitive frequencies area, where the highest pitch corresponds to the top position of image. More importantly, compared with the suppressed frequency response of the peripheral regions of images, the central regions enjoy larger frequency range and are accordingly given more attention. In contrast, the exponential function adopted in the primary setting takes no consideration of the auditory perception characteristics of humans. The translated sound of most image areas are suppressed in the low-frequency area of below 2K Hz, which neither focus on the sensitive frequencies area nor highlight the central regions of images. Hence, such function could be not an appropriate choice.

To effectively evaluate the proposed encoding schemes, amounts of evaluation tests should be conducted with participants. Accordingly, if more modified schemes are provided, much more training and testing efforts will be required, which could go beyond what we can support. Hence, we choose to focus on these two modifications w.r.t. audio length and position-frequency function, as well as the primary one.

4.1. Machine Assessment

The quality of generated modality depends on the quality of the other encoded modality in the cross-modal generative model [38, 30, 6]. Hence, it can be expected to evaluate different encoding schemes by comparing the generated images. In this section, the proposed CBM is chosen as the evaluation reference, as the encoding scheme directly impacts the quality of translated sound for the audio discriminator, then further affects the performance of visual generator, as shown in Fig. 4. However, the adopted MNIST and CIFAR-10 dataset suffer from absent real object or quite complex background. These weaknesses make it quite hard to effectively evaluate the practical effects of different encoding schemes. Therefore, we choose the *Columbia Ob-*
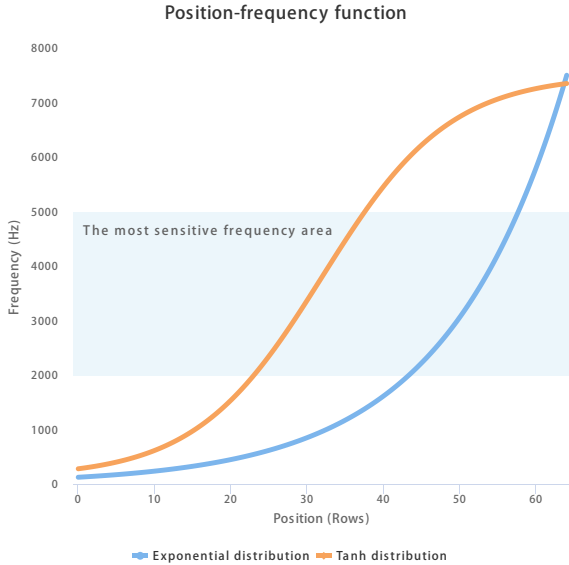

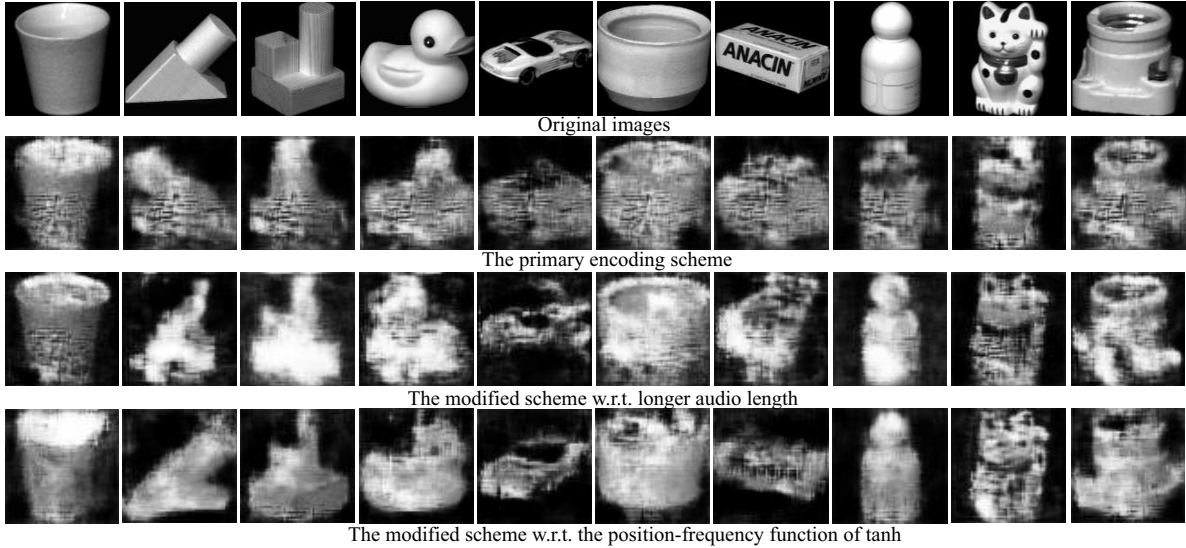Figure 6. Different position-frequency functions of the vOICe translation.

Figure 7. Comparison among the generated image examples using our congenitally-blind model in terms of different encoding schemes.

ject Image Library (COIL-20) [23] as the evaluation dataset, which consists of 1,440 gray-scale images that belong to 20 objects. These images are taken by placing the objects in the frame center in front of a clean black background, meanwhile a fixed rotation of 5 degree around the objects is also performed, which leads to 72 images per object. Considering that this dataset will be also adopted by human-based assessment, we select 10 object categories from COIL-20 for efficiency (i.e., COIL-10), where the testing set consists of 10 selected images of each object, and the remaining ones constitute the training set.

Results. To comprehensively evaluate the generated images, qualitative and quantitative evaluation are both considered. As shown in Fig. 7, in general outlook, as well as in matters of detail, the images of the modified schemes are superior to the primary ones. Obviously, the images generated by the primary scheme suffer from horizontal texture noise. This is because most image area are suppressed in the low-frequency domain, which makes the audio network difficult to achieve effective embeddings for visual generation. By contrast, longer audio track or more effective PF function can contribute to settle such issue. In addition, compared with longer audio signal, the improvement attainable with the proposed PF function of tanh is more significant, such as the details of toy car and fortune cat. And such superiority comes from the effective frequency representation of pixel positions.

For the quantitative evaluation, we choose to compute inception score [31] and human-based evaluation, as shown in Fig. 8. Concretely, we ask 18 participants to grade the quality of generated images from 1 to 5 for each scheme, which correspond to {*beyond recognition, rough outline, clear outline, certain detail, legible detail*}. Then, mean value and standard deviation are computed for comparison. In Fig. 8, it is clear that the qualitative and quantitative evaluation show consistent results. In particular, both of the inception score and human evaluation show that the modified encoding scheme of PF function enjoys the largest improvements, which further confirms the significance of the projection function. And the enlarged audio length indeed helps to refine the primary scheme.

4.2. Cognitive Assessment

The blind participants' feedback or task performance is usually served as the indicator to the quality of encoding schemes in the conventional assessment [3, 33]. Following the conventional evaluation strategy [33], 9 participants are randomly divided into three groups (3 participants per group). Each group corresponds to one of the three encod-
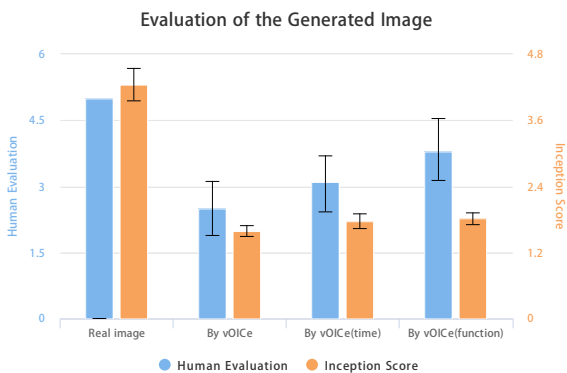

Figure 8. Evaluation of the generated images by our CBM, where different encoding schemes are compared in human evaluation and inception score.
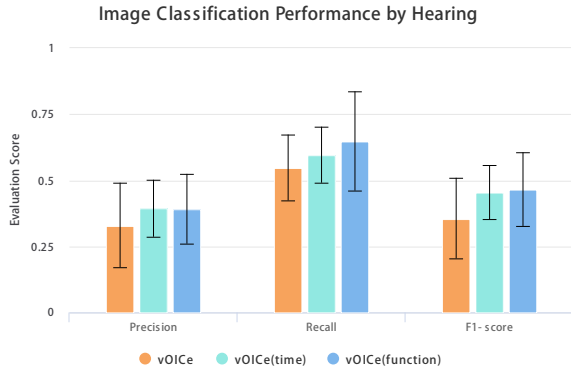
Figure 9. Human image classification performance by hearing the translated sound via different encoding schemes.

ing schemes. The entire evaluation process takes about 11 hours. Before the formal evaluation, the participants are firstly asked to complete the preliminary training lessons to be familiar with the translation rules and simple visual concepts. The preliminary lessons include: identification of simple shape, i.e., triangle, square, and circle; recognition of complex shape, e.g., a normal "L", an upside-down "L", a backward "L", and a backward and upside-down "L"; perception of orientation, e.g., straight white line of fixed-length in different rotation angles; estimation of lengths, e.g., horizonal white line with different lengths; localization of objects, i.e., circles in different places of images. During training, the assistant of each participant plays the pre-translated sound for them, then tell them the concrete visual content within the corresponding image[7]. After finishing amounts of repetitive preliminary lessons, the participants are scheduled to achieve advanced training of recognizing real objects. Concretely, the COIL-10 dataset is adopted for training and testing the participants, where the training procedure is the same as the preliminary lessons. Note that the evaluation test is conducted after finishing the training of each object category instead of all the categories. Finally, the evaluation results are viewed as the quality estimation of the encoded sound and used as the reference for machine-based assessment.

Results. Fig. 9 shows the evaluation results in precision, recall and F1-score w.r.t. different encoding schemes. Specifically, the scheme with modified PF function performed significantly better than the primary scheme ($p < 0.005$, with Bonferroni multiple comparisons correction) in precision, and the recall bars indicate that the scheme with modified PF function performed significantly better than the primary scheme ($p < 0.01$, with Bonferroni multiple comparisons correction). Similar results are observed on F1-score. According to the conventional criterion [33], it can be concluded that the introduced modifications indeed help to improve

---

[7]The training details and image samples can be found in material.

the quality of cross-modal translation. And it further confirms the assumptions about audio length and the characteristics of human hearing. Nevertheless, there still remains a large disparity of classification performance between normal visual and indirect auditory perception. Hence, more effective encoding schemes are expected for the blind community.

Correlation Coefficient	Machine (IS)	Machine (Eva.)
Human (Recall)	0.947	1.000
Human (Precision)	0.952	0.805
Human (F1-score)	0.989	0.889

Table 1. The comparison analysis between machine- and human-based assessment in terms of correlation coefficient.

4.3. Assessment Analysis

The main motivation of designing the cross-modal perception model is to liberate the participants from the boring and inefficient human assessments. According to the shown results above, we can find that the modified scheme of PF function gets the best performance while the primary scheme is the worst one on both assessments. Further, quantitative comparison is also provided in Table 1. Obviously, both assessments have reached a consensus in terms of correlation coefficient, especially the column of *Inception Score* (IS), which confirms the validity of machine-based assessments to some extent.

5. Conclusion

In this paper, we propose a novel and effective machine-assisted evaluation approach to the visual-to-auditory SS scheme. Compared with the conventional human-based assessments, the machine fashion performs more efficiently and conveniently. On the other hand, this paper gives a new direction for the auto-evaluation of SS devices through limited comparisons, more possibilities should be explored in the future, including more optimization schemes and more effective machine evaluation model. Further, the evaluation should be combined with derivable encoding module to constitute a completely automated solver of the encoding scheme without any human intervention, just like seeking high-quality models via AutoML.

Acknowledgments

References

[1] 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 neuroscience*, 10(6):687, 2007. 1, 2

[2] P. Arno, A. G. De Volder, A. Vanlierde, M.-C. Wanet-Defalque, E. Streel, A. Robert, S. Sanabria-Bohórquez, and C. Veraart. Occipital activation by pattern recognition in the early blind using auditory substitution for vision. *Neuroimage*, 13(4):632–645, 2001. 2

[3] P. Bach-y Rita and S. W. Kercel. Sensory substitution and the human–machine interface. *Trends in cognitive sciences*, 7(12):541–546, 2003. 7

[4] D. Bavelier and H. J. Neville. Cross-modal plasticity: where and how? *Nature Reviews Neuroscience*, 3(6):443, 2002. 1, 2

[5] D. Brown, T. Macpherson, and J. Ward. Seeing with sound? exploring different characteristics of a visual-to-auditory sensory substitution device. *Perception*, 40(9):1120–1135, 2011. 2

[6] L. Chen, S. Srivastava, Z. Duan, and C. Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357. ACM, 2017. 3, 6

[7] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 3

[8] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik. Emnist: an extension of mnist to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017. 4

[9] L. G. Cohen, P. Celnik, A. Pascual-Leone, B. Corwell, L. Faiz, J. Dambrosia, M. Honda, N. Sadato, C. Gerloff, M. D. Catala, et al. Functional relevance of cross-modal plasticity in blind humans. *Nature*, 389(6647):180, 1997. 2

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009. 4

[11] S. A. Gelfand. *Essentials of audiology*. Thieme New York, 2001. 6

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[13] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017. 3

[14] N. P. Holmes and C. Spence. Multisensory integration: space, time and superadditivity. *Current Biology*, 15(18):R762–R764, 2005. 2

[15] D. Hu, F. Nie, and X. Li. Deep binary reconstruction for cross-modal hashing. *IEEE Transactions on Multimedia*, 2018. 6

[16] T. Ifukube, T. Sasaki, and C. Peng. A blind mobility aid modeled after echolocation of bats. *IEEE Transactions on biomedical engineering*, 38(5):461–465, 1991. 2

[17] A. J. Kolarik, M. A. Timmis, S. Cirstea, and S. Pardhan. Sensory substitution information informs locomotor adjustments when walking through apertures. *Experimental brain research*, 232(3):975–984, 2014. 2

[18] Á. Kristjánsson, A. Moldoveanu, Ó. I. Jóhannesson, O. Balan, S. Spagnol, V. V. Valgeirsdóttir, and R. Unnthorsson. Designing sensory-substitution devices: Principles, pitfalls and potential 1. *Restorative neurology and neuroscience*, 34(5):769–787, 2016. 1, 2, 3, 4, 6

[19] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 4, 5

[20] Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 3, 4, 5

[21] L. B. Merabet, L. Battelli, S. Obretenova, S. Maguire, P. Meijer, and A. Pascual-Leone. Functional recruitment of visual cortex for sound encoded object identification in the blind. *Neuroreport*, 20(2):132, 2009. 2

[22] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[23] S. A. Nene, S. K. Nayar, H. Murase, et al. Columbia object image library (coil-20). 1996. 7

[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011. 3

[25] U. Noppeney. The effects of visual deprivation on functional and structural organization of the human brain. *Neuroscience & Biobehavioral Reviews*, 31(8):1169–1180, 2007. 2

[26] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016. 5

[27] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413, 2016. 3

[28] C. Poirier, A. G. De Volder, and C. Scheiber. What neuroimaging tells us about sensory substitution. *Neuroscience & Biobehavioral Reviews*, 31(7):1064–1070, 2007. 1, 2

[29] M. Ptito, S. M. Moesgaard, A. Gjedde, and R. Kupers. Cross-modal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind. *Brain*, 128(3):606–614, 2005. 2

[30] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016. 3, 6

[31] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 7

[32] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012. 3

[33] N. R. Stiles and S. Shimojo. Auditory sensory substitution is intuitive and automatic with texture stimuli. *Scientific reports*, 5:15628, 2015. 1, 2, 3, 4, 5, 6, 7, 8

[34] E. Striem-Amit, L. Cohen, S. Dehaene, and A. Amedi. Reading with sounds: sensory substitution selectively activates the visual word form area in the blind. *Neuron*, 76(3):640–652, 2012. 2

[35] B. Thylefors, A. Negrel, R. Pararajasegaram, and K. Dadzie. Global data on blindness. *Bulletin of the world health organization*, 73(1):115, 1995. 1

[36] J. Ward and T. Wright. Sensory substitution as an artificially acquired synaesthesia. *Neuroscience & Biobehavioral Reviews*, 41:26–35, 2014. 1, 2, 4

[37] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint*, 2017. 3

[38] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg. Visual to sound: Generating natural sound for videos in the wild. *arXiv preprint arXiv:1712.01393*, 2017. 3, 6