Supplementary Material

1. Network Setting and Training

1.1. Late-blind Model

In this section, we provide the architectural details of the proposed late-blind model. First, the audio ConvNet follows the VGGish architecture proposed in [4], which achieves excellent results on audio classification task. Second, the visual generator and discriminator mostly adopt the same networks in DCGAN framework [8], except that the out_channel size of last deconvolution layer is set to 1 in MNIST experiments. Third, due to different complexities of MNIST and CIFAR-10 data, we use different visual classifiers for those two datasets. Specifically, a small ConvNet with two convolution layers followed by two fully-connected layers is utilized to classify MNIST digits, and the ResNet-18 [3] is employed as the visual classifier for CIFAR-10 dataset.

The training procedure of the whole late-blind model is composed of three steps. First, the audio ConvNet is pre-trained on MNIST/CIFAR-10 dataset for audio classification, where the audio is obtained by transforming the images with vOICe. And the visual classifier is pre-trained on EMNIST/ImageNet dataset for image classification. Second, the adversarial training strategy in [8] is used to train visual generator and discriminator on EMNIST/ImageNet dataset for visual knowledge learning. Finally, the audio ConvNet, visual generator, and visual classifier are concatenated for cross-modal generation. We firstly fix the visual generator and visual classifier, and fine-tune the audio ConvNet with image classification loss for several epochs. Then we train the visual generator and audio ConvNet together with fixed visual classifier. To be specific, we use a small initial learning rate of 0.001 with Adam optimizer [5] for fine-tuning the audio ConvNet, which decreases by $\frac{1}{10}$ when train the visual generator and audio ConvNet together.

1.2. Congenitally-blind Model

The proposed congenitally-blind model consists of one sound perception module and one cross-modal generation module. For the former, the off-the-shelf large-scale audio classification network of VGGish [4] is employed, but the embedding_dim of the second FC layer is set to 128 and the out_dim is set to the number of classes, i.e., 10. To effectively train such sound model, we set batch_size to 100 and choose the Adam optimizer with learning rate of 0.0002 and beta_1 of 0.5. For the latter, we propose a variant *Auxiliary Classifier GAN* (ACGAN) [7], where the input conditional label is replaced with audio embeddings. More importantly, different from the unimodal processing of ACGAN, the generator deals with visual information while the discriminator focuses on the audio messages. Concretely, the visual generator firstly projects and reshapes the input audio embeddings and noise into certain image shapes (e.g., $8 \times 8 \times 128$ for the MNIST dataset) via one *Fully Connected* (FC) layer and one reshape layer, which is then processed by 3 up-sampling module. Each up-sampling layer is followed by one convolutional layer, as well as batch normalization and ReLU activation. The last layer projects the generated samples into single channel images (in gray scale), where sigmoid function is adopted for activation. The audio discriminator is developed based on the VGGish network, where the activation function of convolutional layers is replaced with Leaky ReLU (with 0.2 alpha) and the discrimination layer and softmax layer are directly performed over the last Flatten layer. The entire cross-modal generation model is optimized via the Adam optimizer (learning rate is set to 0.00002 and beta_1 is 0.5), and the batch_size is set to 100. Moreover, the derivable vOICe translation is derived from the official encoding scheme, and we refactor the official code into a computational graphs for the derivable purpose.

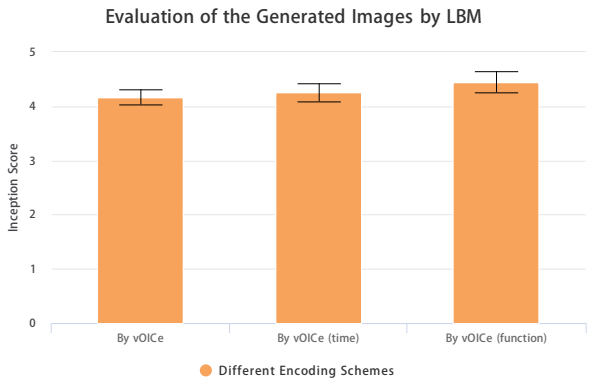


Figure 1. Inception scores of the generated images by our LBM with different encoding schemes.

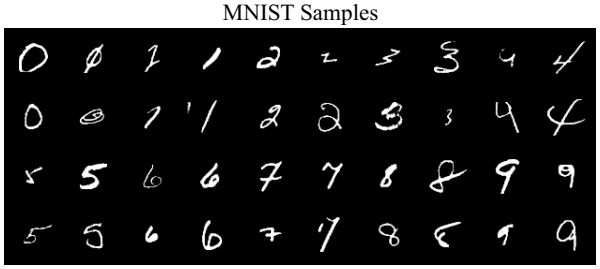
2. Encoding Scheme Evaluation by LBM

In this section, we evaluate different encoding schemes quantitatively and qualitatively. As shown in Fig. 1, we firstly compute inception score of intermediate generated images by our LBM with different encoding schemes. And the results show similar improvements, i.e., the modified encoding scheme of PF function achieves the largest improvements, with quantitative evaluation by CBM and human-based evaluation, which indicates that our LBM framework can also be used for machine-based assessments to some extent. For qualitative evaluation, we show more generated image examples using our late-blind model with different encoding schemes in Fig. 3. Generally, the images of the modified schemes are better than the primary ones in most classes on MNIST/CIFAR-10 datasets. The generated digit images using the modified encoding scheme show more clean background than the primary ones, especially in number 2, 3, 4, and 6. In addition, compared to longer audio length, the proposed PF function of tanh achieves more significant improvement in almost every class, which agrees with the quantitative results in Fig. 1. As for CIFAR-10 dataset, images of the modified schemes contain more detail information, such as windows in airplane images, legs in dog images and meadows in horse images. Meanwhile, there is no obvious improvements in several difficult classes, like automobile, ship, and truck, which confirms the difficulty of realistic objects generation. This is because the LBM learns visual generator and visual classifier from EMNIST/ImageNet datasets and it's hard to transfer learned knowledge to MNIST/CIFAR-10 datasets in cross modal generation, especially when there are extra more classes in ImageNet dataset. Moreover, as for failure case, the modified schemes obtain worse results in number 8, bird, and frog classes, and the reason behind this could be the trained LBM tend to capture detail structure of objects, resulting in overall object contour missing.

3. Dataset Examples

In this section, we show some digit examples of MNIST [6] and EMNIST [1] in Fig. 2. The MNIST dataset is a subset of a much larger dataset of NIST Special Database 19 [2], while EMNIST is a extended MNIST dataset and a variant of the full NIST dataset. Hence, the EMNIST Digits enjoy an increased variability (e.g., size, style, rotation, etc.) and are more challenging [1]. In the handwritten digits generation task, the EMNIST Digits are adopted to provide abundant digit samples for training the visual models of LBM, while the MNIST dataset is used for training the cross-modal perception model.

MNIST Samples



EMNIST Samples

Figure 2. Some samples in the MNIST and EMNIST Digits dataset.

4. Cognitive Evaluation Details

4.1. The preliminary training lessons

In this section, we briefly introduce the preliminary training lessons employed for training the participants. The used images are shown in Fig. 4, and the corresponding sounds can be found in the "examples" folder.

The first lesson. This lesson focused on the initial identification of simple shapes, i.e., circle, triangle, and square. The assistant of each participant randomly selected and played one translated sound for the participant. After playing one sound, the participants were told the concrete content of corresponding image. During the whole first training lesson, each sound (with image) should be played for 15 times. Hence, all the translated sounds were played for 45 times totally. After finishing this lesson, the participants should take a rest for 5 minutes.

The second lesson. Based on the first lesson, the second lesson aimed at the recognition of more complex shapes, i.e., a normal "L", an upside-down "L", a backward "L", and a backward and upside-down "L" (i.e., 7). The assistants randomly selected and played translated sounds for each participant. After playing each sound, the assistants told the participants the concrete shape of corresponding image. In the second lesson, each sound should be played for 15 times totally. Hence, all the translated sounds were played for 60 times. After this lesson, the participants should take a rest for 5 minutes.

The third lesson. In the third lesson, we aimed at the perception of orientation, i.e., straight white bar of fixed-length at 0, 22, -22, 45, -45, or 90 degrees relative to vertical (The positive angles correspond to clockwise rotations). The assistants randomly selected and played translated sounds for

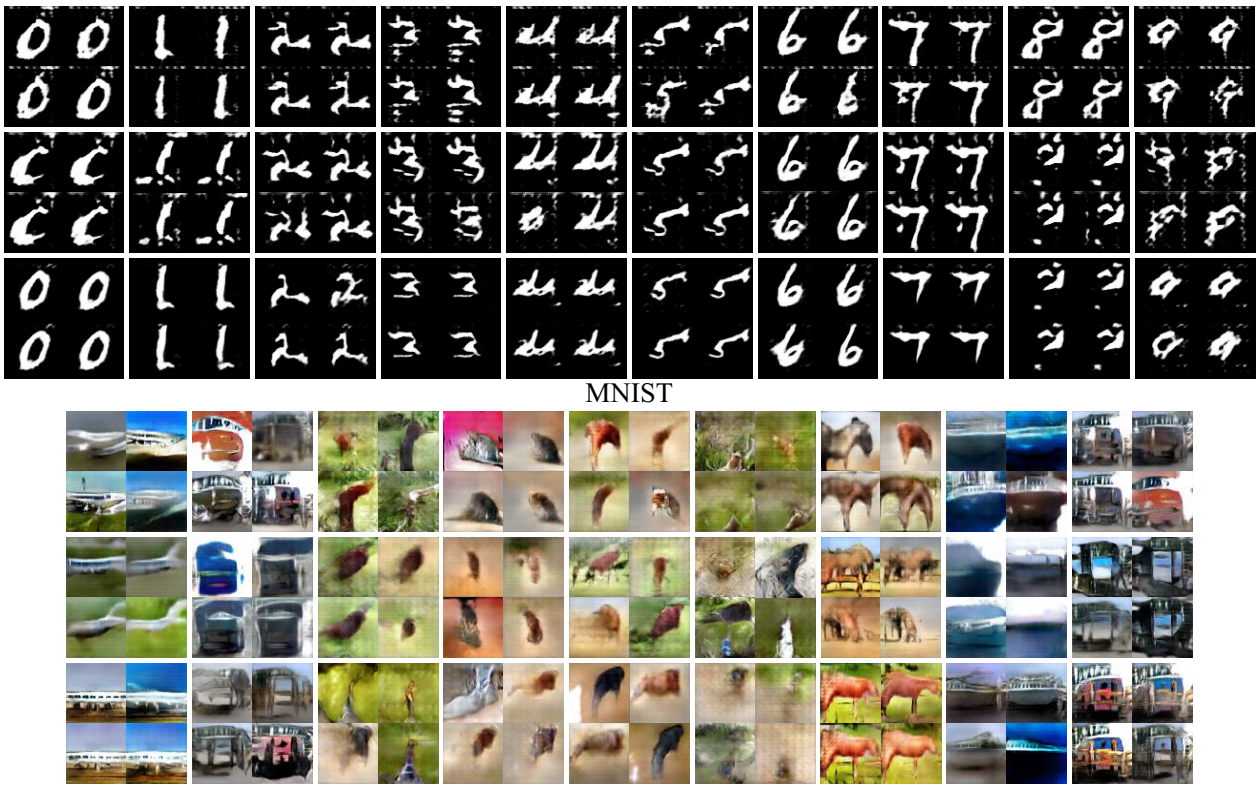


Figure 3. Comparison among the generated image examples of MNIST/CIFAR-10 dataset using our late-blind model in terms of different encoding schemes. For each dataset, the first row represents the primary encoding scheme, the second row represents the modified scheme w.r.t. longer audio length, and third row for the modified scheme w.r.t. the position-frequency function of tanh.

each participant. After playing each sound, the participants were told the concrete orientation of corresponding bar. In the third lesson, each sound should be played for 15 times totally. Hence, all the translated sounds were played for 90 times. After this lesson, all the participants should take a rest for 5 minutes.

The fourth lesson. This lesson focused on the estimation of different lengths, i.e., five bars with different lengths. To improve the sensitivity of lengths, these five bars were also placed in one of four orientations, i.e., 0. 90, 45, and -45 degrees as in the third lesson. During training, the assistants randomly selected and played translated sounds for each participant. After playing each sound, the assistants told the participants the concrete length of corresponding bar (by touch). The translated sound of each image should be played for 15 times totally. Hence, all the sounds were played for 75 times. After finishing this lesson, the participants should take a rest for 5 minutes.

The fifth lesson. In the last lesson, the participants were trained to possess the localization ability, where circles in different places of images (i.e., upper-left, upper-right, bottom-left, bottom-right, and center) were considered. During training, the assistants first randomly selected and played translated sounds for each participant. After playing each sound, the participants were told the position of corresponding circle. The translated sound of each image should be played for 15 times totally. Hence, all the sounds were played for 75 times.

4.2. The advanced training lessons

In the advanced training lessons, all the participants were asked to perform the image classification task by hearing the translated sounds, where the COIL-10 dataset was employed for training and testing. Concretely, the COIL-10 dataset consisted of 10 real objects, such as toy car, fortune cat, bottle, etc. In the training set, each category had 70 image-sound pairs. Before training the participants, they should be told that the images of each object were taken from different angles. Note that the evaluation test was conducted after finishing the training of each object category instead of all the categories. During training, the assistant played the translated sounds of each object for each participant in a certain order. After playing each sound, the assistant told the participant the concrete object and corre-

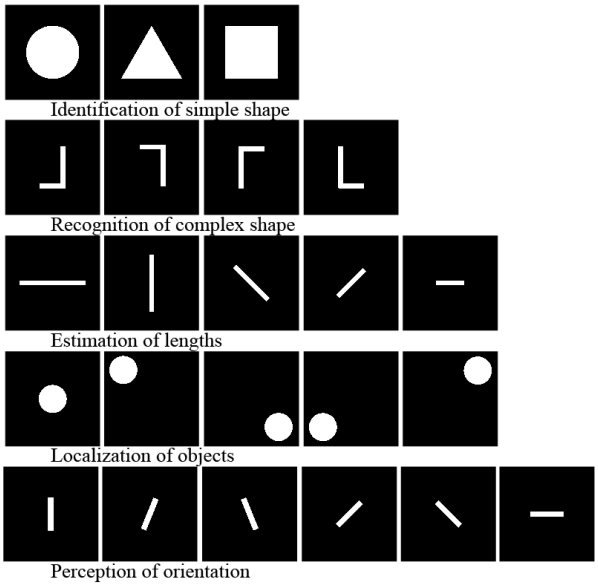


Figure 4. The images used for training the participants in the preliminary training lessons.

sponding angle. In the testing process, 100 sounds (of 10 objects) in the testing set were successively played for the participants, then the participants were asked to answer if the played sound corresponded to the object in the training process. After training and testing all the 10 objects, we evaluated the classification performance in terms of recall, precision, and F1-score.

References

[1] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik. Emnist: an extension of mnist to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017. 2

[2] P. J. Grother. Nist special database 19. *Handprinted forms and characters database, National Institute of Standards and Technology*, 1995. 2

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] 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. 1

[5] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 2

[7] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016. 1

[8] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1