Acoustic-to-Articulatory Inversion by Generative Artificial Intelligence and Fine-Tuning Techniques Using Generative Adversarial Networks

Mingyu Lu[1] and Tatsuya Kitamura[2]

[1]Graduate School of Natural Science, Konan University
[2] Faculty of Intelligence and Informatics, Konan University
Okamoto, Higashinada, Kobe 658-8501, JAPAN

Abstract. In this study, we adopted a fine-tuning technique for the facial video generation method wav2lip to generate magnetic resonance imaging (MRI) articulatory motion movies. We fine-tuned the wav2lip network using a 15-hour dataset of Japanese real-time MRI (rtMRI) movies measured during speech. Our results confirmed that generating rtMRI movies from speech sounds is possible. The structural similarity and normalized mean square error evaluation results indicated good performance. The proposed method will lead to new applications of generative artificial intelligence technology in language learning.

Keywords: GANs, fine-tuning, generative AI, real-time MRI, video generation.

1 Introduction

Recently, the generation of photographs, audio, and videos has advanced considerably, owing to the development of generative artificial intelligence (AI) technologies. Among these, we focused on projects such as wav2lip [1] and sadtalker [2], which can produce lip-synchronous movies from speech. These networks were developed to generate facial motion; however, there are also have some projects for articulatory motion estimation. Tamás Gábor Csapó [3] discusses a method for converting acoustic speech signals into articulatory movements using real-time magnetic resonance imaging (rtMRI) movies and CNN-LSTM of the vocal tract.

In this study, we investigated the application of these techniques to rtMRI movies. We used generative adversarial networks (GANs) to perform the inverse estimation of rtMRI movies from speech because GANs are effective at generating realistic objects and can learn large amounts of data without labels. The results of this study confirmed a method for the inverse estimation of articulatory movements from speech. We believe this method may also be helpful in language learning and assisting people with speech disorders.

2 Method

2.1 Dataset

The dataset used in this study comprised approximately 15 h of rtMRI movies of the head and neck regions in the midsagittal plane, while 29 native Japanese speakers produced Japanese sentences [5]. Four speakers produced 503 sentences in the ATR phoneme-balanced corpus [6], while the others produced Japanese syllables with a career sentence. rtMRI data were measured using a Siemens MAGNETOM Prisma 3T at the ATR Brain Activity Imaging Center in Kyoto, Japan. The size and thickness of the MR scans were 256 × 256 pixels and 10 mm, respectively. Most movies were recorded at 14 fps, while the remainder were recorded at 24 fps. Each speaker's voice was recorded separately using an optical microphone at a sampling frequency of 48 kHz and a resolution of 16 bits. Noise elimination during the MRI was performed using the FRCRN method proposed by Zhao et al. [7]

Fig. 1. Original (left) and trimmed video frame image (right)

The movies were trimmed such that the tongue region was centered, as shown in **Fig. 2**, and the resolution was increased to 1920 × 1080 pixels. The frame rate of the 14-fps movies was upsampled to 24 fps, and the movies split into 5-s segments. **Table 1** lists the data used for training, validation, and testing of the proposed network. The dataset was selected randomly. Finally, every frame and speech sound of the movie was saved in JPEG and WAV files to make it suitable for network input.

Table 1. All-dataset.

Set	Numbers
Training	11,206
Validation	1,401
Test	1,316

2.2 Training of neural networks

In this study, we performed fine-tuning using a network structure published by the lip-video generation project wav2lip [1]. It uses a GAN with a generator and two discriminators, which first extracts speech features (MFCC) from the audio and lip features from the video frames, and finally combines these features to generate a lip movement

video. The training data used in the project, which are available on GitHub, were based on the LRS2 [4] dataset. The training was performed by wav2lip using the fine-tuning method of adding the edited MRI data to the trained checkpoints.

To investigate the training effect, three groups (A, B, and C) of observation experiments were conducted: group A trained the model typically; group B trained the model under the condition that no data from sentences A01 to A12 in the ATR 503 sentences were used during training; and group C trained the model under the condition that no data from speaker HT among the speakers of the ATR 503 sentences were used during training. After training, each group generates ten videos with untrained data.

3 Results and discussion

Figure 2 shows an example of the results generated by the model of the group A. The left side is the original, and the right side is generated after training. The two images appear to be considerably similar.

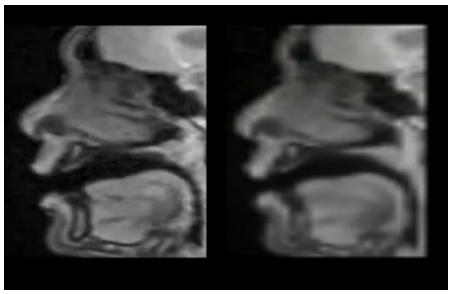

Fig. 2. Original (left) and generated video frame image (right)

The structural similarity (SSIM) and normalized mean squared error (NMSE) for the generated videos were calculated. An SSIM value close to one indicated that the images were similar. By contrast, a smaller NMSE value suggests that the images are dissimilar. **Table 2** presents the results of each group. The SSIM and NMSE are the averages of each group. From this table, it can be concluded that the generated behavior is almost identical to that of the original.

Table 2. SSIM and NMSE for each group.

Group	SSMI	NMSE
A	0.9488	0.008083
B	0.9466	0.006160
C	0.9374	0.007267

Based on the calculation results, we can conclude that the videos are almost identical because the NMSE values of all three groups are small. The SSIM values of groups A and B were almost the same, whereas that of group C was slightly lower. The results showed that the inclusion of a speaker in the training data affected the effectiveness of video generation for that speaker.

4 Conclusion

In this study, the face-generation network was fine-tuned, and rtMRI speech videos were successfully generated. These results open new possibilities for the development and application of deep learning neural networks. For future tasks, we believe that it is important to evaluate the generated videos with even greater precision. In addition, we will consider adding data from other languages, as this study only used Japanese data.

Acknowledgments

This study was supported by MEXT/JSPS Kakenhi (nos. JP20H01265 and JP23K00071).

References

1. K R Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C V Jawahar: A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. Proceedings of the 28th ACM International Conference on Multimedia, pp. 484–492 (2020)
2. Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang: SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8652–8661 (2022)
3. Tamás Gábor Csapó: Speaker dependent acoustic-to-articulatory inversion using real-time MRI of the vocal tract. Proceedings of INTERSPEECH 2020, pp. 3720–3724 (2020)
4. The Oxford-BBC Lip Reading Sentences 2 (LRS2) Dataset, https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html
5. Kikuo Maekawa, Kenya Nishikawa, Takuya Asai, Yukiko Nota, Shinobu Masaki, Yasuhiro Shimada, Hironori Takemoto, Tatsuya Kitamura, Yoshio Saito, Takayuki Kagomiya, Yuichi Ishimoto, Hideaki Kikuchi, Masako Fujimoto, and Yutaka Yagi: Design of Real-Time MRI Articulatory Movement Database. Proceedings of Language Resources Workshop 2020 (2020) (written in Japanese)
6. https://research.nii.ac.jp/src/ATR503.html
7. Shengkui Zhao, Bin Ma, Karn N. Watcharasupat, and Woon-Seng Gan: FRCRN: Boosting Feature Representation using Frequency Recurrence for Monaural Speech Enhancement. Proceedings of ICASSP 2022, pp. 9281–9285 (2022)