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

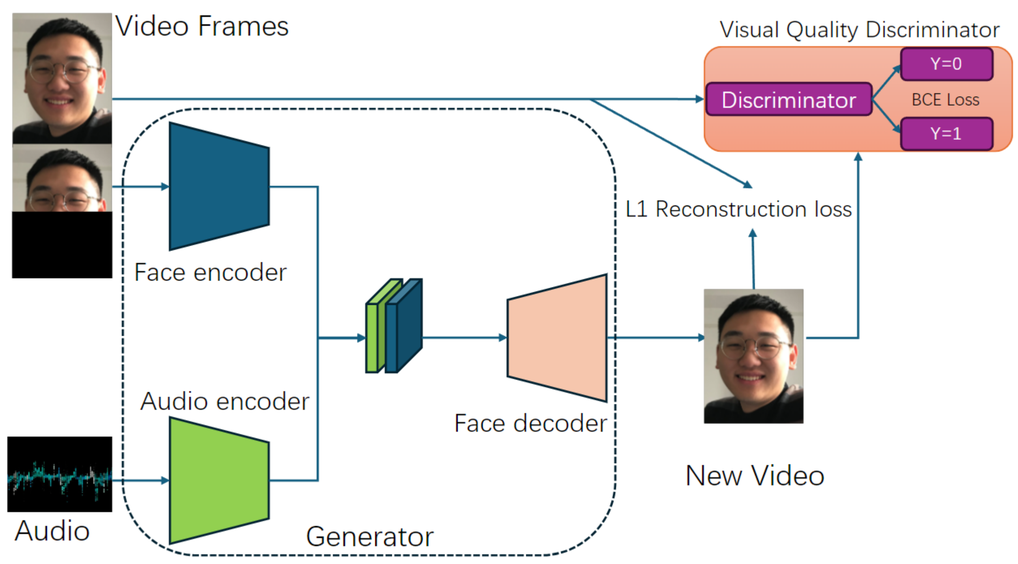
Introduction

We adopted a **fine-tuning** technique for the facial video generation method wav2lip [1] to generate magnetic resonance imaging (MRI) articulatory motion movies.

Method

● Network

By using the wav2lip network which opened at 2020. 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.



● Dataset

The 15-hour dataset of Japanese real-time MRI (rtMRI) movies measured during speech [2] is used for training. And the rtMRI movies were measured using Siemens MAGNETOM Prisma 3T.

Table 1. Parameters for rtMRI movies.

Size	256 x 256 px
Thickness	10 mm
Fps	14 fps / 27 fps

Because of the machine for measure rtMRI could not using for voice. We used an optical microphone to take the voice.

Table 2. Parameters for voices.

Sampling frequency	48 kHz
Resolution	16 bits

Noise elimination during the MRI was performed using the FRCRN method proposed by Zhao et al. [3]

● Edit

The movies were trimmed such that the tongue region was centered, as shown in Fig. 1

Table 3. Parameters for edited movies.

Size	1920 x 1080 px
Sampling frequency	24 fps
Times	5 s

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

Table 4. All-dataset.

Set	Numbers
Training	11206
Validation	1401
Test	1316

Training

● Training

Performed fine-tuning using the dataset we edited and the wav2lip network which is trained for lip-video generation.

Result

The training effect was investigated in three groups..

- **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
- **group C**:trained the model under the condition that no data from speaker HT among the speakers of the ATR 503 sentences were used

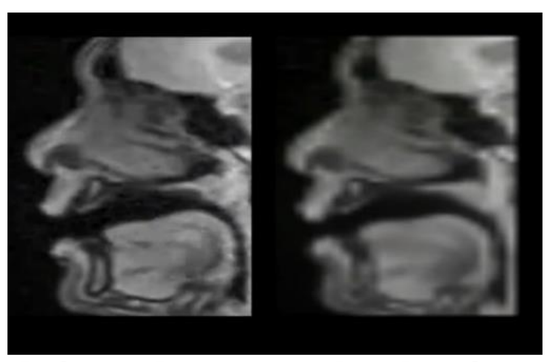


Fig. 2. Original (left) and generated video frame image (right)

Table 5. SSIM and NMSE for each group.

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

By calculated the SSIM and NMSE to show the effect. Based on the results, NMSE values of all group are almost identical. And on the SSIM values **Group C** is the slightly lower that A and B.

It showed that the inclusion of a speaker in the training data affected the effectiveness of video generation for that speaker.

Conclusion

In this study, the face-generation network was fine-tuned, and rtMRI speech videos were successfully generated. And for **future tasks**:

I. evaluate the generated videos with even greater precision
II. 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] Prajwal et al., Proc. 28th ACM International Conference on Multimedia, 484–492 (2020).
[2] Maekawa, Acoustical Science and Technology, (2024).
[3] Zhao et al., Proc. ICASSP 2022, 9281-9285 (2022).