Data Descriptor

Korean Audio-Visual Dataset of Characters in 3D Animation: Construction and Validation

Soobin Hyun , Yeongmin Son , Jae Wan Park *

Posted Date: 9 October 2023

doi: 10.20944/preprints202310.0514.v1

Keywords: anime character; 3D animation; audio-visual dataset

Data Descriptor

Korean Audio-Visual Dataset of Characters in 3D Animation: Construction and Validation

Soobin Hyun [1], Yeongmin Son [2] and Jae Wan Park [3,*]

[1] Global School of Media, Soongsil University, 50, Sadang-ro, Dongjak-gu, Seoul 07027, Republic of Korea; ana9686@soongsil.ac.kr
[2] Department of Media, Soongsil University, 50, Sadang-ro, Dongjak-gu, Seoul 07027, Republic of Korea; son342@soongsil.ac.kr
[3] Global School of Media, Soongsil University, 50, Sadang-ro, Dongjak-gu, Seoul 07027, Republic of Korea; jaewan.park@ssu.ac.kr
* Correspondence: jaewan.park@ssu.ac.kr

Abstract: Characters are one of the most important elements in composing digital animation. The appearance and voice of a character should be designed to express the personality and values of the character. However, it is not easy for animation producers to harmoniously match the appearance and voice of a character. Advances in deep learning technology have made it possible to overcome this limitation. To achieve this, firstly, an audio-visual dataset of characters is required. In this study, we construct and verify a Korean audio-visual dataset consisting of frontal face images of various characters and short voice clips. We developed an application that can automatically extract the frontal face image and a short voice clip of a character by collecting videos uploaded to YouTube. Through this, a dataset consisting of a total of 1,522 face images and a total of 7,999 seconds of voice clips was built based on 490 characters. Furthermore, we automatically label characters by gender and age to validate the dataset. The dataset built in this study is expected to be used in various deep learning fields, such as classification, generative adversarial networks, and speech synthesis.

Dataset: https://github.com/Dripmaster/Audio-Visual_3D_Animation_Dataset

Keywords: anime character; 3D animation; audio-visual dataset

1. Summary

Demand for digital media continues to increase worldwide with the advancement of information technology. Especially, the size of the digital animation market is growing remarkably. Characters are one of the most important elements in composing digital animation. The appearance and voice of a character should be designed to encompass the characteristics of the character so that the audience can quickly grasp the personality and values of the character. Appearance and voice of a well-designed character can help viewers understand the characteristics of the character [1]. Specifically, matching the appearance and voice of a character is important. If the appearance and voice of characters are inconsistent, it will lower the interest of the viewers in the animation. However, it is not easy for animation producers to find a voice that matches the character of the animation. Looking for voice actors that match the character is a time-consuming task. Recently, owing to the rapid development of deep learning technology, a technology for generating a human face using a voice [2,3] or, on the contrary, using a face is being studied [4]. Furthermore, a dataset for this study has been established [2]. However, there is no dataset consisting of the appearance and voice of an anime character.

Therefore, in this study, we aim to construct and verify a Korean audio-visual dataset of characters. However, data collection is not easy because of the exaggerated nature of the character's face and the limited number of animations. Therefore, in this study, we have two research questions:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2023 doi:10.20944/preprints202310.0514.v1

how can we effectively build a dataset, and what information can we additionally supply to the audio-visual dataset of characters for various experiments and verification?

This study is conducted as an extension of basic research for the Korean audio-visual dataset of characters [5]. We collected 3D animation videos uploaded to YouTube. For a more convenient collection, we found a channel that only uploaded animations and crawled all animations in that channel. Then, to effectively collect data, we build a Mediapipe-based application to extract frontal facial images and voices from the collected videos [6]. Furthermore, additional information, such as gender and age, was added to the character to support various experiments. Through this, a high-quality dataset consisting of a total of 1,522 frontal face images and a total of 7,999 seconds of audio clips was finally built. The dataset of this study is expected to be used to apply artificial intelligence in various animation-related industries.

2. Data Description

This dataset consists of 490 animation characters, and each character consists of 1–3 audio clips of at least 3 s and 1–5 images. The total number of audio clips is 1,317 and total time is 7,999 s. The total number of images of the front face of the character is 1,522. Table 1 lists the specifications of our audio-visual dataset and Figure 1 shows a part of the dataset loaded into Python.

Table 1. Specifications of the Audio-visual Dataset.

	Characters	Frontal Face Images	Audio Clips
Total Number	490	1522	1317
Average Number of Each Character	-	≈ 3.1	≈ 2.68
Average Length of Audio Clips	-	-	≈ 16.3s
Size	-	128 * 128(Width x Height)	7999s(Total Length)

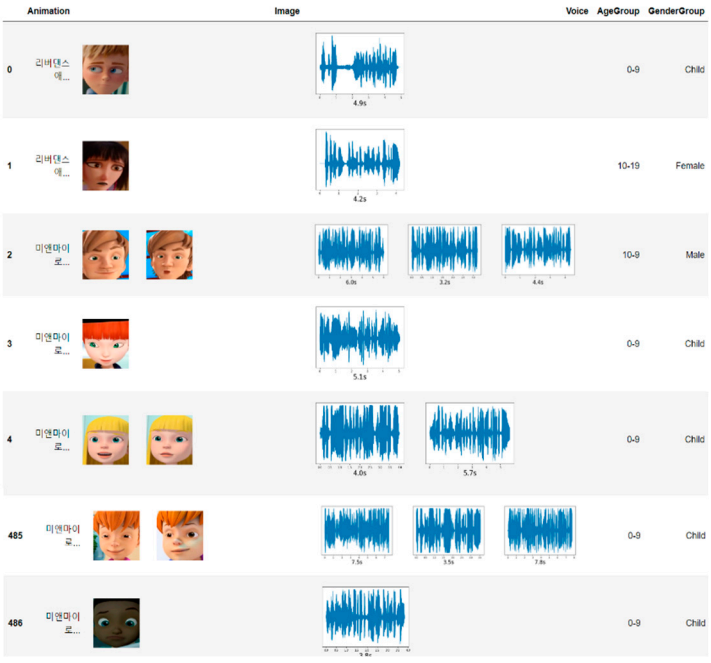


Figure 1. Part of the Dataset Loaded into Python.

The length of the audio clips in the dataset varies from 3–19s. In other words, as shown in Figure 2, the minimum length of audio clips is 3s, and the maximum length is 19s. Generally, 4s is considered as a high percentage.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2023 doi:10.20944/preprints202310.0514.v1

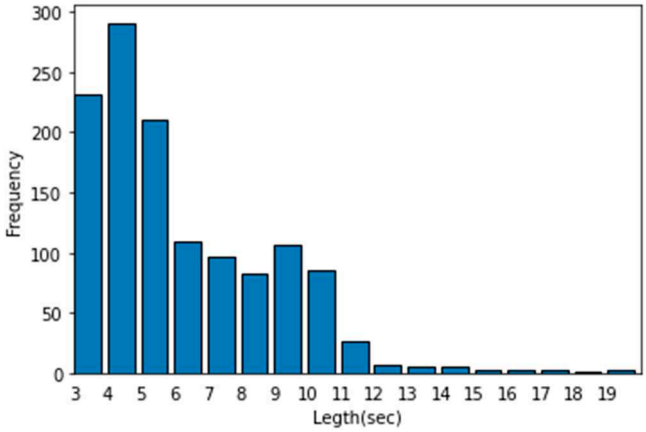


Figure 2. Frequency According to the Audio Clip Length.

The gender group consisted of males, females, and children with 3.4, 34.2, and 62.4%, respectively(Figure 3), and the age groups are divided into 0–9, 10–19, and over 20, and the proportions are 40, 44.7, and 15.3%, respectively(Figure 4). In this dataset, there is a bias in the number of characters according to gender and age group. This was because of the lack of 3D animations dubbed in Korean that were uploaded to YouTube. It is determined that additional data collection using movie streaming sites other than YouTube is necessary.

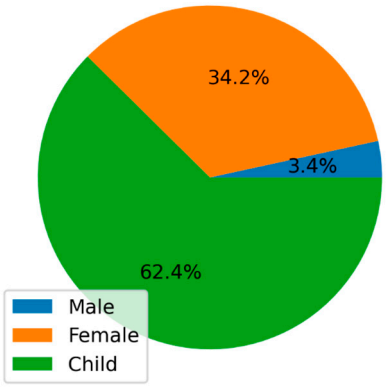


Figure 3. Character's Male to Female Ratio.

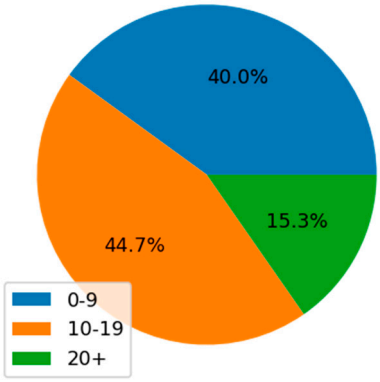


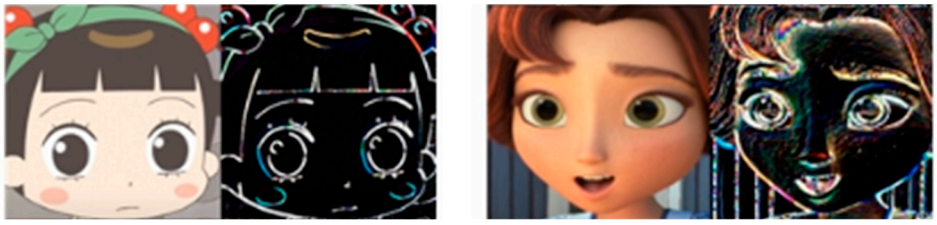
Figure 4. Character's Age Group Ratio.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2023 doi:10.20944/preprints202310.0514.v1

3. Methods

3.1. Data Collection

To construct this dataset, 3d animations uploaded on YouTube were first collected. Channels that upload only animation were collected, and all videos of the channel were crawled to minimize manual work. 2D characters have limitations in using them as training data because exaggerated features appear differently for each animation. Therefore, only 3D characters were used. A contour-detection processing technique was used to automatically distinguish 2D and 3D animation from the collected videos. 3D images have unclear and soft borders owing to contrast, whereas 2D character borders were clear and only areas where the values around pixels change abruptly were detected, resulting in a high black-and-white ratio. Therefore, among the various kernels, the Scharr filter, which expresses the boundary of the 3D image most indistinctly and smoothly [7], was used (Figure 5). A threshold value was set through the ratio of color to black and white, and the videos were classified as 2D if this value exceeded and 3D if not.



(**a**) 2D Character (**b**) 3D Character

Figure 5. Application of the Scharr filter.

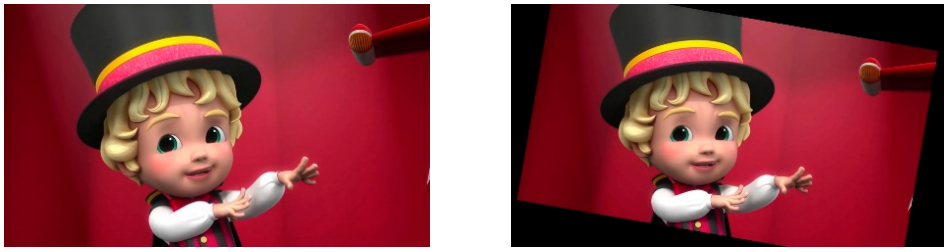
After obtaining only the videos featuring the 3D character, the face image and voice of the character were extracted based on the application developed using Mediapipe Face Detection. Face Detection is a high-performance face recognition AP I that recognizes faces through six landmarks [8]. First, the section where the character's face appears is extracted using face detection. Then, based on the extracted section, the screen of the section is captured and saved, and the audio is saved in the wav file format [9].

The following rules were applied to obtain a section in which a face appears. First, images that are too far from the screen are not saved because they are of poor quality. For this, we set the model index of MODEL_SELECTION, a property of face detection, to 0 so that the small face in a distance is not recognized. The model index can be set from 0 to 1 and is suitable for recognizing a person within 2 m and a person within 5 m from the monitor, respectively. Images that are too far from the screen will not be of good quality. Second, when several people appear on one screen, the face of the largest character on the screen is captured. In general, when several people appear on one screen, the fact that the speaker usually appears large on the screen was applied. Third, as an animal with a human facial structure is also recognized as a human, only images with a probability of 80% or more of being a face are considered. Fourth, the section with a video playback time of less than 2 s is deleted. This is because the section where the character appears for less than 2 s is usually a scene showing facial expressions without exclamations or lines; therefore, it is not suitable for use as a training data. Using these rules, we were able to effectively extract face images and voices.

3.2. Data Preprocessing

The images of the character were saved one by one every 0.5 seconds, and only the most frontal and expressionless images were extracted for ease of learning. For the rotated image of the expressionless face, the degree of rotation centered on both eyes was first obtained, rotated so that it was parallel, and then the face was cut and stored (Figure 6).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2023 doi:10.20944/preprints202310.0514.v1



(**a**) Before Rotation (**b**) After Rotation

Figure 6. Image Parallel Rotation.

Furthermore, to facilitate learning, we increased the quality of data by deleting unnecessary images as follows: characters with more than half-closed eyes, non-frontal characters, characters wearing glasses, and characters with a beard (Figure 7).

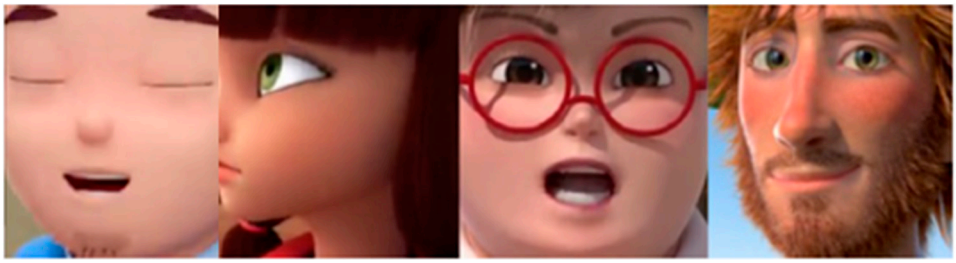
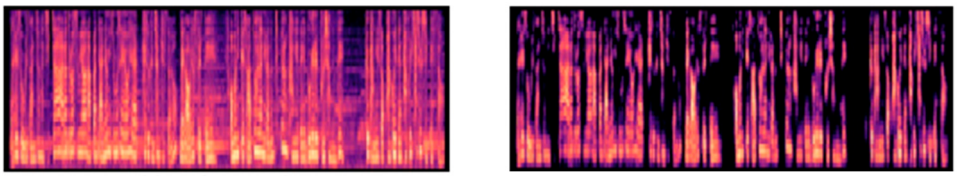


Figure 7. Image Data Preprocessing.

In animation, background music (BGM) frequently appears in addition to the voice of the character. Therefore, BGM or sound effects are removed through noise cancellation (Figure 8). Additionally, using an open-source voice detection (VAD) algorithm, the silent section of the audio was removed [10,11].



(**a**) Raw Data (**b**) Noise Reduction

Figure 8. Audio Preprocessing.

3.3. Additional Information of Characters

For various experiments, verification, and ease of learning, that is, to build a high-quality dataset, we labeled each character with gender and age. Gender and age classification were automatically labeled using Naver Clova Face Recognition and an open API, and unrecognized images were removed [12]. For face images that are too young to be classified as gender, the API outputs 'children' instead of gender. Because the age predictions of the API are provided in the form of ranges, we used the median of each prediction range to label age groups. Therefore, our dataset was categorized and labeled into gender groups and age groups.

4. Verification

To verify the validity of the dataset built in this study, we present the results of using character images as training data for gender and age classification. For this, EfficientNet was used for transfer learning. EfficientNet has high accuracy in image classification and high-learning efficiency in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2023 doi:10.20944/preprints202310.0514.v1

transferfor gender and age classification. For this, EfficientNet was used for transfer learning. EfficientNet has high accuracy in image classification and high-learning efficiency in transfer learning [13]. Fine-tuning was performed using two variations: EfficientNet-B0 and EfficientNet-B3. The structures of EfficientNet-B0 and EfficientNet-B3 are listed in Tables 2 and 3, respectively.

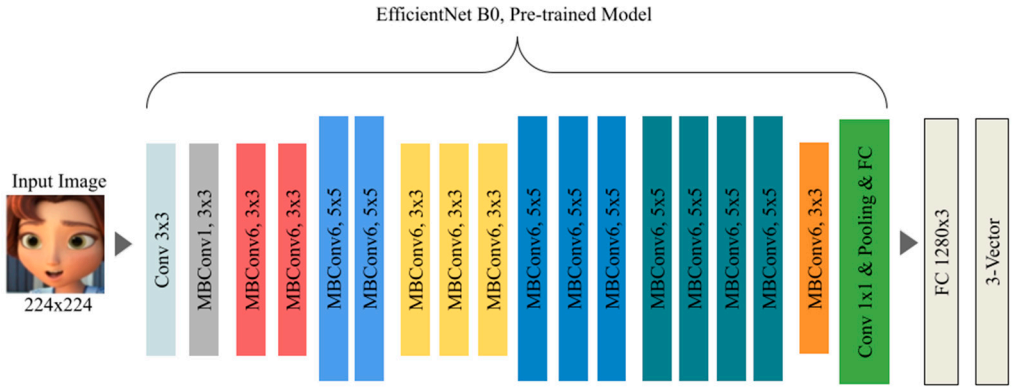
Table 2. EfficientNet-B0 Architecture.

Stage	1	2	3	4	5	6	7	8	9
Operator	Conv3x3	MBConv1, k3x3	MBConv6, k3x3	MBConv6, k5x5	MBConv6, k3x3	MBConv6, k5x5	MBConv6, k5x5	MBConv6, k3x3	Conv1x1 &Pooling &FC
Resolution	224x224	112x112	112x112	56x56	28x28	14x14	14x14	7x7	7x7
#Channels	32	16	24	40	80	112	192	320	1280
#Layers	1	1	2	2	3	3	4	1	1

Table 3. EfficientNet-B3 Architecture.

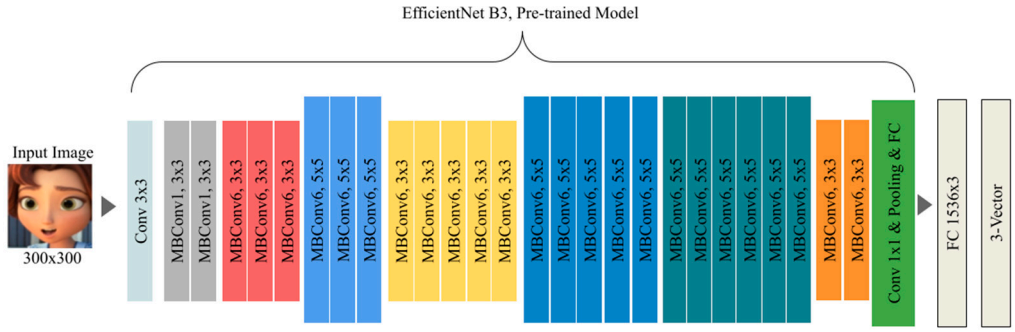
Stage	1	2	3	4	5	6	7	8	9
Operator	Conv3x3	MBConv1, k3x3	MBConv6, k3x3	MBConv6, k5x5	MBConv6, k3x3	MBConv6, k5x5	MBConv6, k5x5	MBConv6, k3x3	Conv1x1 &Pooling &FC
Resolution	224x224	112x112	112x112	56x56	28x28	14x14	14x14	7x7	7x7
#Channels	40	24	32	48	96	136	232	384	1536
#Layers	1	2	3	3	5	5	6	2	1

To use the dataset constructed in this study as training data, each image was resized to 224 x 224 and 300 x 300 according to the structure of EfficientNet-B0 and EfficientNet-B3, respectively [13]. In addition, FC layers for age and gender classification were added to the last FC layer of each model. Xavier initialization was used for adding FC layers [14,15]. Figure 9 shows the size of the added FC layer.



(**a**) Fine-tuning with EfficientNet-B0

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 October 2023 doi:10.20944/preprints202310.0514.v1



(**b**) Fine-tuning with EfficientNet-B3

Figure 9. Fine-tuning Model

The two deep learning models showed 78.4% and 81% accuracy in gender classification and 81.6% and 80.3% accuracy in age classification, respectively (Table 4). Figure 10 shows the change in loss according to epoch. From 30 epochs onwards, no further learning has progressed. This is because of the lack of data.

Table 4. Training Result for Each Model.

	EfficientNet-B0		**EfficientNet-B3**	
Category	Gender	Age	Gender	Age
Training Time(s)	15m 8s	12m 59s	16m 56s	16m 25s
Number of Classes	3	3	3	3
Accuracy(%)	78.4	81.6	81.0	80.3

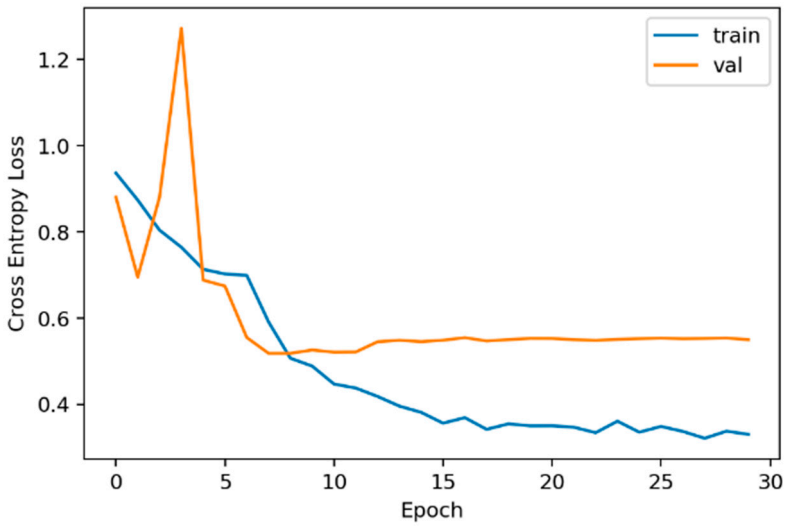


Figure 10. Graph of loss according to epoch.

5. Conclusion

The aim of this study was to construct and verify a Korean audio-visual dataset of characters in 3D animation. To do this, we first crawled the animation data uploaded to YouTube. Then using MediaPipe, we developed an application for extracting the frontal face images and voices of the character. Utilizing this application, we were able to effectively build a high-quality dataset.

The dataset built in this study comprises 490 characters, the total number of audio clips is 1,317, and the total number of face images of characters is 1,522. Each character consisted of an average of 16.3 s of voice clips and an average of three or more face images. Moreover, each character is labeled with gender and age automatically for verifying this dataset. The two deep learning models built

with transfer learning for dataset validation showed an accuracy of up to 81% in gender classification and 81.6% in age classification. The loss of accuracy is owing to the lack of data.

Additional data will be collected in future studies. Furthermore, we will utilize this dataset to build a model that can generate a character by inputting a voice clip and a model that generates a voice by inputting a character. We also plan to release this dataset for free use by researchers and students. The dataset built in this study is expected to be used to apply artificial intelligence in various animation-related industries.

Author Contributions: Conceptualization and methodology, J.W.P.; software, S.H. and Y.S.; validation, Y.S.; writing—original draft preparation, S.H. and Y.S.; writing—review and editing, J.W.P.; supervision, J.W.P.; funding acquisition, J.W.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. NRF-2021R1F1A1063884).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study, and any future updates, are openly. available at https://github.com/Dripmaster/Audio-Visual_3D_Animation_Dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hahm, S., Convergence Research on the Speaker's voice Perceived by Listener, and Suggestions for Future Research Application. *IJASC* **2022**, 11, 55-63.
2. Oh, T. H.; Dekel, T.; Kim, C.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; Matusik, W., Speech2face: Learning the Face behind a Voice. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, California, USA, 16 June 2019; pp.7539-7548.
3. Wen, Y.; Raj, B.; Singh, R., Face Reconstruction from Voice Using Generative Adversarial Networks. *NeurIPS* **2019**, 32.
4. Lu, H. H.; Weng, S. E.; Yen, Y. F.; Shuai, H. H.; Cheng, W. H., Face-based Voice Conversion: Learning the Voice behind a Face. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20-24 October 2021; pp.496-505.
5. Hyun, S.; Son, Y.; Park, J. W., Building a Korean Audio-Visual Dataset of Characters in 3D Animation. In Proceedings of the 10th O2O International Symposium on Advanced and Applied Convergence, Seongnam, Korea, 17-19 November 2022; pp.121-126.
6. MediaPipe. https://mediapipe.dev/ 25 September 2023.
7. Ciresan, D. C.; Meier, U.; Masci, J.; Gambardella, L. M.; Schmidhuber, J., Flexible, High Performance Convolutional Neural Networks for Image Classification. In Proceedings of Twenty-second International Joint Conference on Artificial Intelligence, Barcelona Catalonia, Spain, 16-22 July 2011.
8. Perisic, H.; Strömqvist, T., PhysiKart: A 2D Racing Game Controlled by Physical Activity through Face-tracking Software, degree of Bachelor, Linköping University, Sweden, **2022**.
9. Nam, K., A Study on Processing of Speech Recognition Korean Words. *JCCT* **2019**, 5, 407-412.
10. Python Interface to the WebRTC Voice Activity Detector, https://github.com/wiseman/py-webrtcvad/ 25 September 2023.
11. An, S. J.; Choi, S. H., Voice Activity Detection Based on SNR and Non-Intrusive Speech Intelligibility Estimation. *IJIBC* **2019**, 11, 26-30.
12. Open API to the Naver CFR, https://developers.naver.com/products/clova/face/ 25 September 2023.
13. Tan, M.; Le, Q., Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, PMLR, California, USA, 9-15 June 2019; pp.6105-6114.
14. Glorot, X.; Bengio, Y., Understanding the Difficulty of Training Deep Feedforward Neural Networks. In Proceedings of the Thirteenth International Conference on artificial intelligence and statistics, PMLR, Sardinia, Italy, 13-15 May 2010; pp.249-256.

 doi:10.20944/preprints202310.0514.v1

15. Kang, M. J.; Kim, H. C., Comparison of Weight Initialization Techniques for Deep Neural Networks. *IJACT* **2019**, 7, 283-288.