STUDY OF VOICE GENERATION METHOD SUITABLE FOR CHARACTERS BASED ON HUMAN COGNITIVE CHARACTERISTICS

Shogo Saito, Sho Ooi, and Mutsuo Sano

[1]Graduate School of Osaka Institute of Technology
[2, 3] Osaka Institute of Technology

ABSTRACT

Previous studies have attempted to estimate existing voices from images of animated characters as a way to generate voices suitable for animated characters, but without good results. Therefore, in this study, to link the voice characteristics to match the animation character with the image, we devised a method to analyze the voice's tendency to not be uncomfortable and then establish the ratio of voice learning data based on the analyzed tendency data. Specifically, this study prepares multiple voices for one illustration of an anime character, asks subjects to evalu- ate the voices, and calculates an evaluation based on the evaluation values. In experiments, we conducted an evaluation experiment using the one-pair comparison method, calculated the distri- bution of learning data based on the evaluation values obtained, and prepared for the subsequent learning process.

KEYWORDS

Synthesized speech, Voice generation, Character

1. INTRODUCTION

Some e-books have a voice reading function so that users can read books while doing something. From this point of view, it can be inferred that the e-book market has been growing steadily in recent years. However, the quality of voice reading is low, and audio comics, in which voice actors voice the characters, are costly. In addition, when comics and audiobooks are converted into video images, there is sometimes a sense of discomfort when voices are added to 2D characters. In this study, we consider whether this discomfort can be resolved using information technology. We believe that the relationship between face and voice, in which the characteristics of the voice are inferred when a person sees a face, can be applied to 2-D characters as well. In this paper, we propose a voice generation model that can be applied to 2D characters by acquiring trend data based on the relationship between faces and voices, using tacotoron2[1] and waveglow[2], and learning voice based on these data. In addition, it is possible to estimate voice from a human face [3] and face from voice [4]. However, what if this voice quality estimation is not for humans but for animated characters? Since the positions of the parts of a human face are fixed to some extent, it is easy to mechanically estimate the voice from the face image. However, in the case of animated characters, the positions of the eyes and nose vary greatly depending on the artist's drawing style, etc. Therefore, when estimating voice quality from animated characters, it is necessary to use an estimation system based on atmospheric recognition rather than a mechanical voice estimation system based on image recognition, etc. However, according to a study[5] that investigated the effect of voice on age estimation, voice has an effect on age

David C. Wyld et al. (Eds): AIBD, MLSC, ACSTY, NATP, CCCIoT, SVC, SOFE, ITCSS -2024
pp. 199-208, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140416

estimation, but in the case of animated characters, the illustration and age do not always match, making estimation difficult in the current situation. In a previous study of this research, we proposed a system for estimating and generating voice from character image features alone, without estimating age or gender from character face images[6]. However, in this study, we extracted the eye parts of a character's face from the image data, and used the extracted eye image data as the main axis of voice learning data, based on a questionnaire in which participants were asked to imagine the gender of the voice and the age of the voice from the image alone, without using voice. As a result, there was a discrepancy between human voice estimation from a character's face and machine voice estimation from the character's face. Therefore, it is considered difficult for machines to estimate the voice quality of animated characters in current research. Therefore, in this study, we considered analyzing the tendency of voices that do not sound strange to the characters in order to map voice and image features that are appropriate to the characters. For this purpose, this experiment involves the generation of synthetic voice data to be learned and the preparation of images to be used in comparison experiments. We also conducted this research and experiment based on the belief that it is possible to generate speech with less discomfort by adjusting the ratio of training data based on the obtained trend data. In a previous study [7], synthetic voice was generated using voice training data, and a comparison experiment between image and voice was conducted using the one-pair comparison method. Therefore, in this study, the speech data used in the pairwise comparison method are randomly selected from a corpus of three different types of speech data. In addition, in order to improve the accuracy of the data, we will conduct the experiment on a larger number of subjects than in the previous study [7] to differentiate the two methods. In this study, after voice learning is performed by allocating training data based on voice acquired from a speech corpus and trend data using multiple types of character illustrations, voice learning using tacotoron2 based on previous research [1] is performed to generate voice as imagined. In this study, we also acquired data such as the ages of the experimental participants in the comparison experiment. Therefore, this paper discusses the allocation of training data based on trend data, paying attention to whether there are differences in the imagined voice depending on the age of the participants. A simplified diagram of the entire flow from the image of an unknown character to the generation of an imagined voice is shown below [Fig. 1].

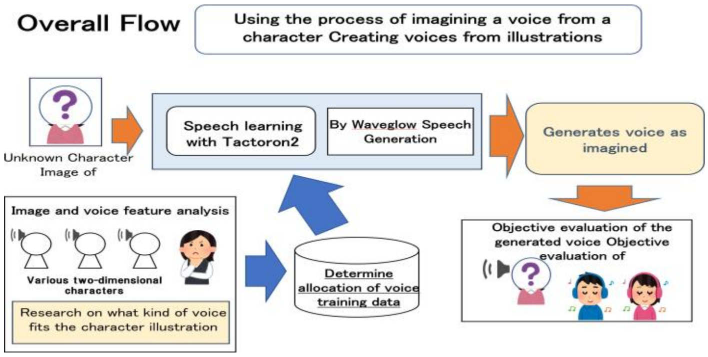

Overall Flow
Using the process of imagining a voice from a character Creating voices from illustrations
Unknown Character Image of
Speech learning with Tactoron2
By Waveglow Speech Generation
Generates voice as imagined
Image and voice feature analysis
Various two-dimensional characters
Research on what kind of voice fits the character illustration
Determine allocation of voice training data
Objective evaluation of the generated voice Objective evaluation of

Fig. 1. Overall flow.

2. Related works

2.1. Speech Estimation Using Eye Images

In a previous study, a survey was conducted to investigate which parts of a character people look at to imagine its voice, and using the results, a study was conducted to examine whether it is possible to perform voice estimation from image features in a simpler manner [6]. In this study, the first step was to investigate which parts of a face image people look at when they see a face image to imagine the voice. As a result of the research, it was found that people imagine voices by looking at ”hair shape,” ”hair color,” and ”eye shape” [8], and among these three points, ”eye shape” was extracted as the most likely to imagine voices, and ”hair shape” as the next most likely [Figure 2]. In a previous study [6], the extracted eye and hair images were then associated with the character voice assigned to the extracted character source, and a classifier and synthetic voice were created. Then, the classifier is used to assign a synthetic voice to an unknown character, aiming to generate a voice that is close to the image. However, in this research, only images are used for voice assignment, so when humans listen to the generated synthetic voice, there is a large possibility that it will not be as the image is perceived. Therefore, this research creates a feature analysis model using character images and voice, and uses the feature model to generate voice, which we believe can generate voice that matches the image better than conventional methods.

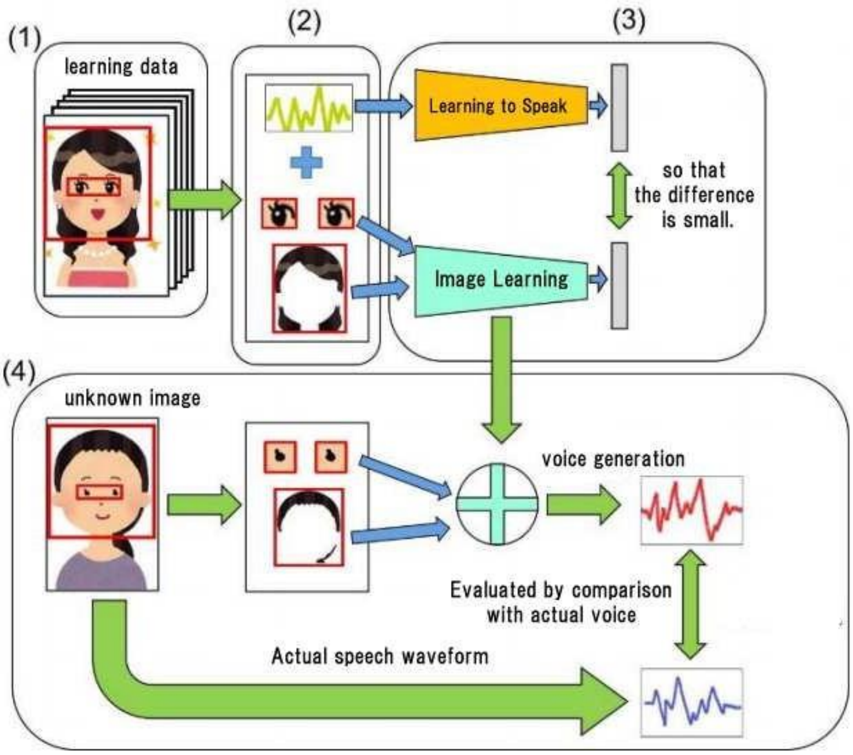


Fig. 2. A method to generate sound from image extraction of eyes and hair.

2.2. Research On Voice Actor Casting

Other studies have focused on the selection of voice actors for characters . [9] In this study, multiple voice data performed by that voice actor to voice a particu- lar character were collected, and acoustic features were extracted from these voice data[Figure 3]. Next, a model was developed to predict the appropriate acoustic features for new characters as well. This was a model that learned the relationship between the acoustic features obtained and the impression values of the character. This proposed a learning approach for voice generation. However, these approaches used general voice data rather than voice data from actual voice actors, which remains a challenge when generating voices for characters such as those in ani- mation. In this study, we focus on generating voices that better match character illustrations by utilizing voice data from voice actors who play characters similar to those in anime, which are used for comparative experiments using the pairwise comparison method, speech synthesis, and voice learning. This will make the voices generated from the character illustrations in this research less uncomfortable.

.

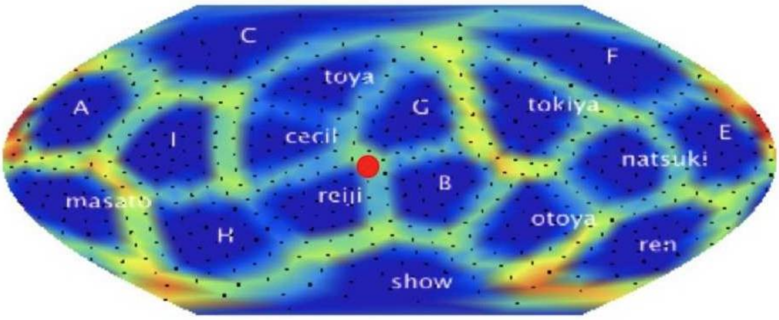


Fig. 3. Learned results of impression values and acoustic features.

2.3. Generation of Voice from Face Image of Person using DNN

A study on the relationship between human face and voice [10] uses machine learn- ing to learn the relationship between face and voice. In this study, machine learning is used to learn the relationship between face and voice using a Deep Neural Net- work (DNN). In addition, the proposed DNN model is a textual dataset. In this study, text data is obtained from VGGFace2, a text dataset, and voice data is ob- tained from VoxCeleb2, a video dataset, and voice learning is performed from face images using data from just under 6000 faces and voices. As a result, the system has succeeded in generating speech captured as rough features. In addition, Figure 4 below shows an image that serves as an overall view of speech generation from human face images using DNN. [Figure 4]

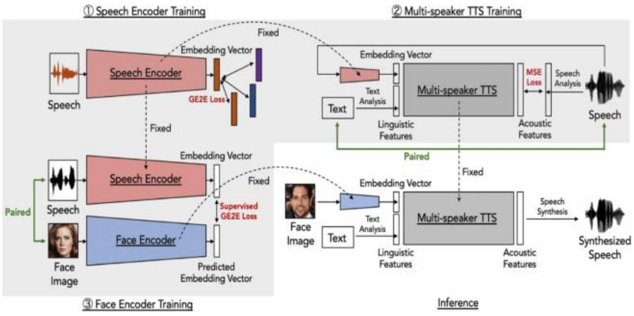


Fig. 4. Voice Generation from Facial Images Using DNN

3. PROPOSED METHODS

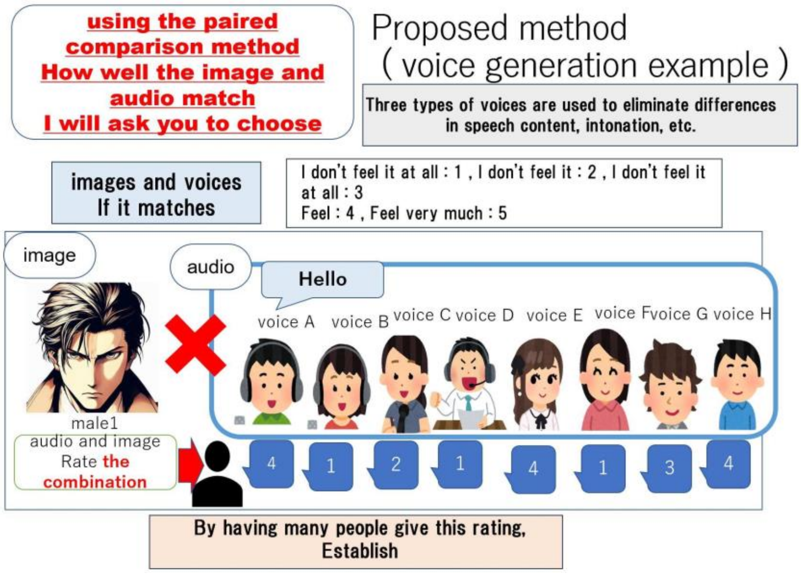
3.1. Trend Data Survey using Pairwise Comparison Method of Image and Sound

The aim of this study is to enhance the accuracy of voice estimation and genera- tion for character images. In pursuit of this goal, this study aims to conduct a more comprehensive exploration of the intricate relationships between character illustra- tions and voices. Previous research [6] has indicated that eye shape, in particular, plays a significant role in voice imagination, and this feature has been applied in voice estimation and learning. However, these approaches relied on visual character impressions and did not delve into the utilization of actual speech sounds.

Therefore, this research proposes that it is possible to generate voice that matches the image of an unknown character by finding an allocation of voice data that matches the character’s illustration and voice, and then learning and generat- ing voice using that allocation. However, there remains the problem of how to find the right allocation of training data to produce the voice that matches the image. To explain briefly, I will conduct a trend data analysis experiment using images and voices of several human subjects. The voice can be created as imagined.

Specifically, numerous participants were asked to assess the degree of agreement between randomly selected voices and character illustrations through a pairwise comparison method. Participants rated the degree of agreement on a 5-point scale (ranging from 1 to 5), where 1 signified no agreement at all, and 5 indicated a high level of agreement. Participants were instructed to rate each combination of image and voice. For instance, if a participant selected the image ”male1” and the voice ”A,” it indicated that they perceived agreement between the image and voice, resulting in a rating of 4, and so forth. By involving a substantial number of

participants in this process, it is believed that the voices generated will align more consistently with the broadly defined character image as perceived by individuals. Figure 5 provides a summarized representation of the trend data generated from the image-voice combination scoring survey.



using the paired comparison method How well the image and audio match I will ask you to choose

Proposed method (voice generation example)

Three types of voices are used to eliminate differences in speech content, intonation, etc.

images and voices If it matches

I don’t feel it at all : 1 , I don’t feel it : 2 , I don’t feel it at all : 3
Feel : 4 , Feel very much : 5

By having many people give this rating, Establish

Fig. 5. Example of voice generation.

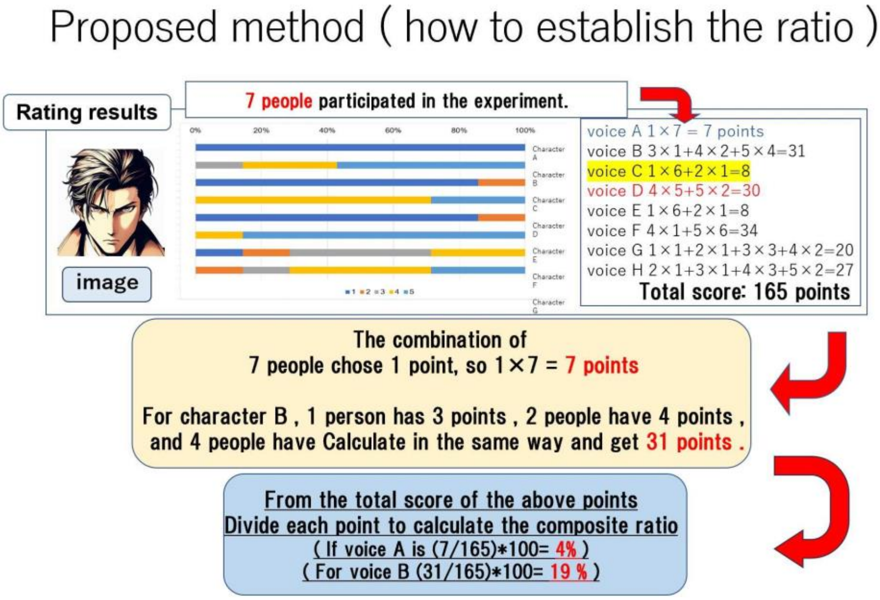
3.2. How do we establish the percentages?

Then, we propose the utilization of the trend data obtained in subsection 3.1 to establish the training data allocation ratio. As depicted in Figure 6 below, let’s consider the participation of seven individuals in this experiment. The results show that 7 people selected a rating of 1 for the combination of image and voice A, 1 person chose 3 for the combination of image and voice B, 2 people chose 4, and 4 people chose 5. Similarly, we can assume the existence of combinations of voice and image spanning from voice C to voice H.

In this hypothetical scenario, one person would select image A and the other person would select image B. Based on these selections, we calculate the sum of points for the combination of image and point A as 1 x 7 = 7 points, and for the combination of image and voice B as 3 x 1 + 4 x 2 + 5 x 4 = 31 points. Similarly, we proceed to calculate the total points for all combinations up to voice

H, which amounts to 165 points in this case. Subsequently, the relative allocation ratio for each voice is computed. For example, for voice A, the allocation would be (7/165)*100 = 4%, and for voice B, it would be (31/165)*100 = 19%, and so forth. This process of percentage calculation is visually represented in

Figure 6, offer- ing an illustrative example of the training data ratio for a male character’s voice.



Proposed method (how to establish the ratio)

Rating results

7 people participated in the experiment.

image

voice A 1×7 = 7 points
voice B 3×1+4×2+5×4=31
voice C 1×6+2×1=8
voice D 4×5+5×2=30
voice E 1×6+2×1=8
voice F 4×1+5×6=34
voice G 1×1+2×1+3×3+4×2=20
voice H 2×1+3×1+4×3+5×2=27
Total score: 165 points

The combination of
7 people chose 1 point, so 1×7 = 7 points

For character B , 1 person has 3 points , 2 people have 4 points , and 4 people have Calculate in the same way and get 31 points .

From the total score of the above points
Divide each point to calculate the composite ratio
(If voice A is (7/165)*100= 4%)
(For voice B (31/165)*100= 19 %)

Fig. 6. Example of Percentage Calculation from Trend Data.

3.3. Determining The Number Of Training Data

Finally, we outline a method for determining the quantity of training data. The number of training data is derived by referencing the previously calculated ratio for each voice to the image. Using the example provided earlier, where the training data for voice A is 4%, for voice B is 19%, for voice C is 5%, and so on, we can assume that there are 100 training data available for each voice. Consequently, we would obtain 19 different voice data sets for voice A, 19 for voice B, 5 for voice C, and so forth. These training data are then employed for voice learning and voice generation, aiming to produce voices that closely resemble the character’s image.

To enhance the authenticity of the generated voices, this study does not cate- gorize voice data by gender. Consequently, there’s a mix of voice data from female and male characters in the learning process. The voice learning process employs tacotron2, while waveglow is used for voice generation. The proposed method is based on the trend data, allocation of training data according to the trend data, and the determination of the quantity of training data. Figure 7 shows a concise visual representation of the proposed method.

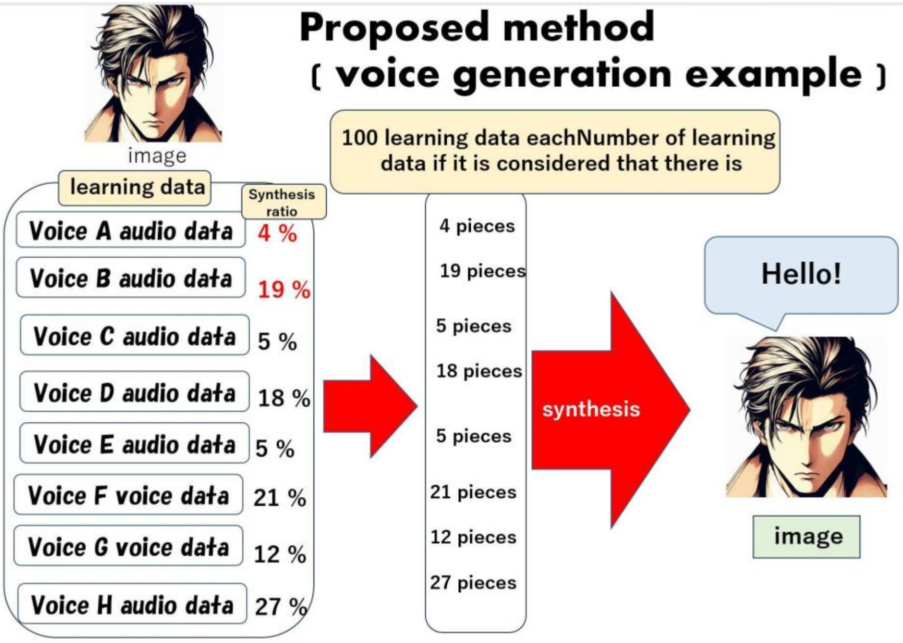


Fig. 7. Example of determining the number of training data.

4. EXPERIMENTS

For the experiment, we utilized 8 types of character illustrations generated by Bing AI as shown in Figure 8. Additionally, we employed 24 types of audio data, with three being randomly selected from the available speech character corpus for each character illustration. In the experiment, we designed a program to present each participant with one image and one sound in 192 different combinations, randomly generated. Participants were instructed to provide a rating from 1 to 5, where 1 sig- nified minimal agreement between the image and sound, 2 indicated low agreement, 3 meant neither agreement nor disagreement, 4 represented high agreement, and 5 denoted very high agreement when the combination of image and sound matched. To explore potential differences in perception based on the participant's age and gender, we requested participants to provide their age and gender during the experiment.

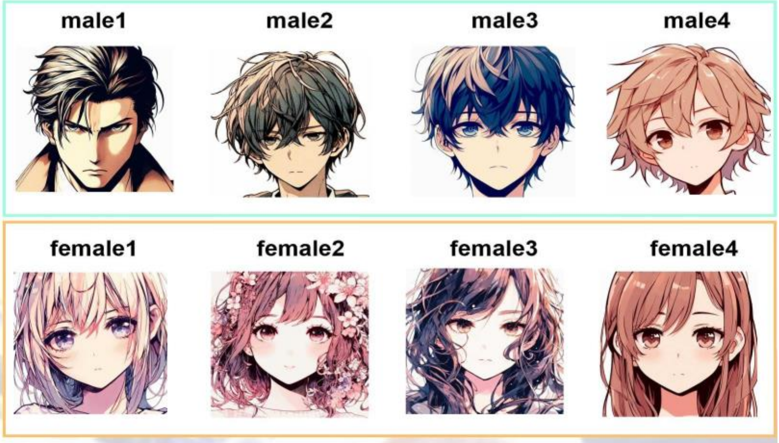


Fig. 8. Character illustrations used.

This study conducted experiments on 10 males between the ages of 10 and 23. We used wireless earphones for the experiment, considering that noise can reduce concentration. This study used a laptop and one display for the display. The partic- ipants were not made conscious of the content of the audio that was played during the experiment, and after a brief explanation was given before the experiment, they were asked to perform the experiment. Figure 9 shows the experiment.

5. Results And Discussions

The results of the experiment using the one-pair comparison method, which involved 10 participants as shown in Table 1. The table in the figure displays the composite ratios based on the trend data gathered. From this table, we observe that for the ”male 1” image, the training data for ”voice 1” was allocated at 6%, ”voice 2” at 6%, ”voice 3” at 20%, and so on. For the ”female 1” image, the training data for ”voice 1” was allocated at 17%, ”voice 2” at 22%, ”voice 3” at 5%, and so on.

These results, represented as percentages, suggest that higher-quality speech synthesis can potentially be achieved by increasing the number of training data while maintaining the same ratios when actually synthesizing speech. Furthermore, the results of the experiment indicate that only the ”male 4” character’s results have an average ratio. This is attributed to ”male 4” having a neutral appearance. As a result, the synthesized voice output for ”male 4” is less likely to cause significant discomfort, based on the experiment’s findings.

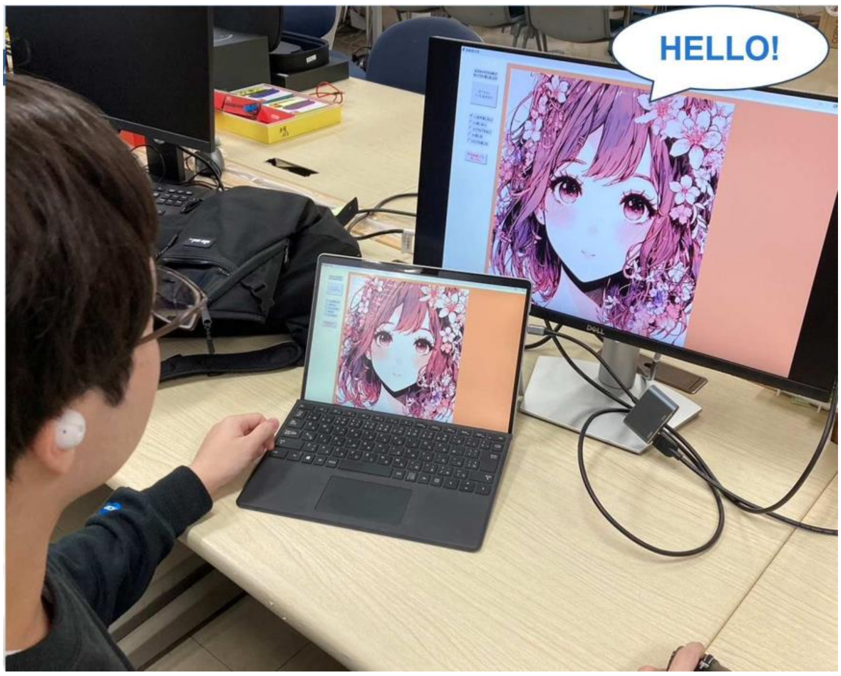


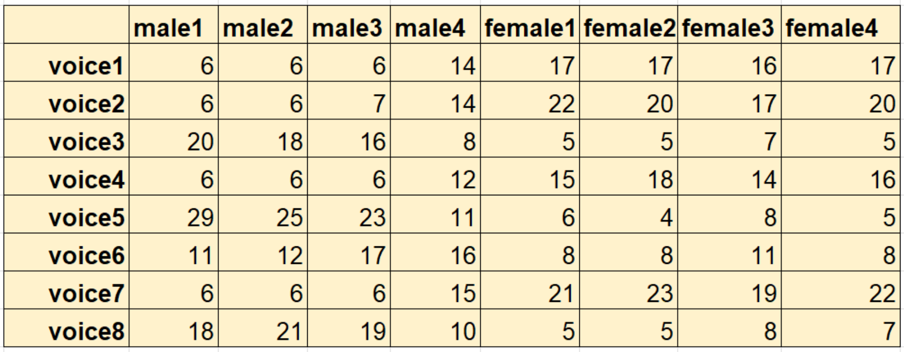
Fig. 9. Experimental Scene.

Age and gender data were collected during the experiment. Considering these demographics, all participants were male and relatively close in age, ranging from 19 to 23 years old. As a result, there were no significant anomalies in the measurement data due to age or gender differences.

6. Conclusion And Future Prospects

This study offers a novel approach by combining various character illustrations and multiple voices, conducting a trend analysis to identify voice characteristics that align with the character, and proposing a method to utilize the obtained trend data for adjusting learning data.

In an experiment involving the cooperation of 10 male participants, we obtained the voice ratios necessary for generating the voices of 10 different characters. As a result, male voices were chosen for the male characters, and female voices for the female characters. However, each voice possessed distinct nuances. By integrating these voices, it is possible to generate voices that do not appear unnatural to human listeners. It was also noted that androgynous characters had an equal mixture of male and female voices, an intriguing observation.



	male1	male2	male3	male4	female1	female2	female3	female4
voice1	6	6	6	14	17	17	16	17
voice2	6	6	7	14	22	20	17	20
voice3	20	18	16	8	5	5	7	5
voice4	6	6	6	12	15	18	14	16
voice5	29	25	23	11	6	4	8	5
voice6	11	12	17	16	8	8	11	8
voice7	6	6	6	15	21	23	19	22
voice8	18	21	19	10	5	5	8	7

Table 1. Percentage of training data obtained from experimental results.

In order to improve the accuracy of the trend data, we would like more people to participate in the experiment, and in order to take into account differences in perception due to gender, we would like to conduct the experiment not only with males but also with females. In the future, we will generate synthetic voices based on the calculated training data ratio, and objectively evaluate whether the generated voices match the images. In addition, future research will also be conducted to generate voices that are more natural to the listener by matching the character's voice to the individual's preferences. This could lead to more sophisticated and personalized voice generation.

References

[1] Shen,Jonathan,et al.”Natural tts synthesis by conditioning wavenet on mel spectrogram pre- dictions.” 2018 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP).IEEE,2018.

[2] Prenger, et al. ”Waveglow: A flow-based generative network for speech synthesis.” IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, 2019.

[3] Smith,Harriet MJ,etal. “Concordant cues in faces and voices: Testing the backup signal hy- pothesis,” Evolutionary Psychology 14.1 (2016): 1474704916630317.

[4] Shunji Kuritsu and Haruka Asano.” Estimation of appearance from sound of unknown objects in still images”, Nisshin 72nd Annual Meeting, 2008

[5] Yoko Kasahara and Yukio Itsukushima, ”Voice Recognition: Influence of Voice Pattern on Age

Estimation", Law and Psychology, 2007, 6, 1, 71-84

[6] Noboru Omichi, Sho Ooi, and Mutsuo Sano, "Investigation of Relationship between 2D Char- acter Features and Voices for Automatic Generation of Audiobooks," Information Processing Society of Japan, Kansai Branch, Section Meeting, 2021.

[7] Shogo Saito, Sho Ohi, and Mutsuo Sano, "A Study of Speech Generation Model for 2D Char- acters," Information Processing Society of Japan, Kansai Branch, Section Meeting, 2022

[8] Noboru Omichi, Sho Ohi, and Mutsuo Sano. "Study on Feature Extraction Method from 2D Character Illustration based on Human's Cognitive Characteristics for Automatic Voice Esti- mation" AICCC'21, (2021).

[9] Erika Sakai, Akinori Ito, Takayuki Ito. "Trend Analysis of Game Characters and Voice Qual- ity," (Visualization, Character Animation, Visual Expression and Art Science Forum 2016).

[10] Technical Report of the Institute of Image Information and Television Engineers 40.11. The Institute of Image Information and Television Engineers, (2016).

[11] Shunsuke Goto, Kotaro Oonishi, Yuki Saito, Kentaro Tachibana, Koichiro Mori. "Multi- Speaker Voice Synthesis Using Predicted Embedding Vectors from Facial Images." Proceedings of the Acoustical Society of Japan Spring Meeting 2020, Lecture Paper Collection, 2-Q-49, pp. 1141–1144, (2020).

AUTHORS

Shogo Saitou was born in Hyogo, Japan, entered the graduate school of Osaka Institute of Technology in 2022. Belongs to Information Processing Society of Japan.

Sho Ooi was born in Osaka, Japan. He received his Ph.D. degree in information science from Osaka Institute of Technology in 2018. He jointed an assistant pro- fessor with the Faculty of Information Science and Engineering in Ritsumeikan University, Japan. Currently, he is an assistant professor with Faculty of Informa- tion Science and Technology in Osaka Institute of Technology, Japan. His research interests include computer vision, cognitive science, pattern recognition, and edu- cation technology. He is a member of the Information Processing Society of Japan (IPSJ), Institute of Electronics, Information, and Communication Engineers (IE- ICE), Institute of Image Electronics Engineers of Japan (IIEEJ), Robot Society of Japan (RSJ), and IEEE.

Mutsuo Sano 1983 Graduated with a degree in Precision Engineering, Graduate School of Engineering, Kyoto University. Joined the Telecommunications Research Laboratory of Nippon Telegraph and Telephone Public Corporation (now NTT) in the same year.Since then, he has been engaged in research on robot vision, im- age recognition, and content management technology.Since 2002, he has been a professor at the School of Information Science and Technology, Osaka Institute of Technology.Since then, he has been engaged in research on robot communication control, cooking support, cooking media, and cognitive rehabilitation.Cognitive re-habilitation.D. in Engineering. D. in Engineering. He is a member of The Institute of Image Information and Television Engineers, The Institute of Image Information and Television Engineers, The Institute of Image Information and Television Engi- neers, The Institute of Artificial Intelligence, The Robotics Society of Japan, and IEEE.