Technology and Health Care 32 (2024) S543–S553
DOI 10.3233/THC-248047
IOS Press

A computer-aid speech rehabilitation system with mirrored video generating

Yang Cao, Chunjiang Fan*, Gang Li, Jian Huang and Jinli Zhang
Department of Rehabilitation Medicine, Rehabilitation Hospital, Wuxi, Jiangsu, China

Abstract.
BACKGROUND: Aphasia is a communication disorder that affects the ability to process and produce language, which severely impacting their lives. Computer-aid exercise rehabilitation has shown to be highly effective for these patients.
OBJECTIVE: In our study, we proposed a speech rehabilitation system with mirrored therapy. The study goal is to construct a effective rehabilitation software for aphasia patients.
METHODS: This system collects patients' facial photos for mirrored video generation and speech synthesis. The visual feedback provided by the mirror creates an engaging and motivating experience for patients. And the evaluation platform employs machine learning technologies for assessing speech similarity.
RESULTS: The sophisticated task-oriented rehabilitation training with mirror therapy is also presented for experiments performing. The performance of three tasks reaches the average scores of 83.9% for vowel exercises, 74.3% for word exercies and 77.8% for sentence training in real time.
CONCLUSIONS: The user-friendly application system allows patients to carry out daily training tasks instructed by the therapists or the prompt information of menu. Our work demonstrated a promising intelligent mirror software system for reading-based aphasia rehabilitation.

Keywords: Aphasia, speech rehabilitation, mirrored therapy, video generation

1. Introduction

Aphasia is a communication disorder that affects the ability to process and produce language [1,2]. It is typically caused by damage to the areas of the brain responsible for language function, such as the left hemisphere in right-handed individuals [3,4,5]. This condition can occur suddenly, often as a result of a stroke or head injury, or it may develop gradually due to progressive neurological conditions [6]. Treatment for aphasia often involves speech and language therapy, which aims to improve communication skills and help individuals find alternative ways to express themselves [7,8].

Aphasia rehabilitation is a specialized form of therapy designed to help individuals with aphasia regain and improve their language abilities following brain injury or damage that has affected their language centers [9,10]. The therapeutic approach may include exercises to strengthen language abilities, techniques to enhance comprehension, and strategies to compensate for specific language deficits. Speech and language therapists, often with expertise in aphasia management, are crucial in guiding the rehabilitation process [11,12]. They employ evidence-based techniques, exercises, and technology to maximize language recovery and communication outcomes.

*Corresponding author: Chunjiang Fan, Department of Rehabilitation Medicine, Rehabilitation Hospital, No.100 Beitang Road, Wuxi, Jiangsu, China. E-mail: fanchunjiang1980@163.com. ORCID: 0009-0004-6185-1068.

Furtherly, several assistive devices are used for the above physical training. Clinically, transcranial magnetic stimulation (TMS) [13,14,15] or transcranial direct current stimulation (tDCS) [16,17] are used for activating the language-related cerebral cortex in the brain. Several studies have shown this technology potentially facilitates the reorganization and plasticity of neural circuits involved in language production and comprehension [18,19,20]. However, owing to the unclear mechanism of neural plasticity, we can't predict the positive effect for all patients. Virtual Reality (VR) technologies are increasingly used in aphasia rehabilitation to create immersive and interactive environments that stimulate language processing and communication skills [21]. While there is a potential risk of accidents or injuries if individuals become disoriented or lose their balance while immersed in a virtual environment. VR technology can be relatively expensive, making it less accessible to individuals and healthcare facilities with limited financial resources. Besides, computer-based software programs specifically designed for language therapy can engage individuals in interactive exercises and games that target different aspects of language recovery, such as vocabulary building, sentence formation, and comprehension [22].

Nowadays, computer-aided devices have been increasingly used in aphasia rehabilitation. This technology provides personalized and interactive rehabilitation exercises for these patients, aiming to improve language and communication skills in individuals with aphasia. Mahmoud et al. proposed an automatic speech recognition platform for aphasia recognition [23]. Agarwal developed a novel computer vision technique for speech therapy [24]. These specialized communication software are designed to assist individuals with aphasia in expressing themselves and understanding others. These software applications often include various features, such as picture-based communication boards, text-to-speech functionality, and word prediction to support speech output and language comprehension.

While these technologies have shown promise in addressing some challenges in aphasia rehabilitation, it is essential to recognize that it is not a replacement for traditional therapy but rather a complementary tool [25]. Human interaction, personalized therapy plans, and skilled guidance from speech and language therapists remain critical components of successful aphasia rehabilitation programs. As the field of VR technology continues to advance, ongoing research and clinical trials will help determine its long-term efficacy and potential integration into mainstream aphasia rehabilitation practices.

Mirror therapy is a rehabilitation technique used to alleviate pain and improve movement in individuals with various neurological conditions, particularly those who have experienced limb injuries or undergone amputations [26]. The fundamental principle behind mirror therapy is based on the concept of mirror neurons in the brain. These mirror neurons are brain cells that fire when we perform a particular action and when we observe someone else performing the same action [27]. By using mirrors to create visual illusions, the brain is tricked into perceiving the reflection of the unaffected limb as the affected one. This visual feedback can lead to significant changes in the brain's perception and motor control. It's important to note that aphasia can present in various forms, and the extent of language impairment depends on the location and extent of brain damage [28,29]. Hence, mirror therapy could be explored for aphasia treatment. We developed a mirror rehabilitation software system with aritificial intelligence technologies firstly.

2. Method

2.1. System design

The architecture of our proposed mirror aphasia rehabilitation system is shown in Fig. 1. It includes image acquisition, speech recording, video generation engine, speech synthesis engine, evaluation

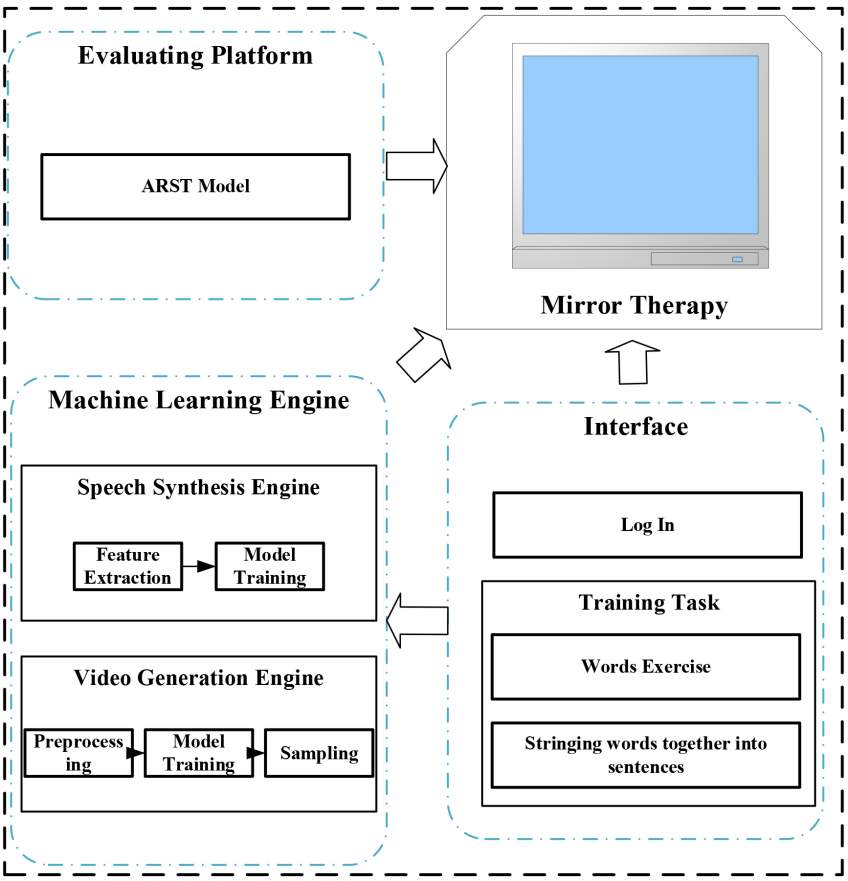


Fig. 1. The architecture of our proposed mirror aphasia rehabilitaton system.

platform and training task modules. The image acquisition and speech recording collects patients' facial photos and speech for video generation and speech synthesis. And these two engines utilized machine learning technologies for generating patients' speaking video as training instructions. The evaluation platform gives the scores assessed by the ARST algorithm, aiding the patients performing the training tasks.

2.1.1. Speech recording

In the next work, the patient must read several sentences in a sequence for audio recording. At the same time, the microphone records the speech, which can be stored as digital audio files for later analysis.

2.1.2. Image acquisition

Before the rehabilitation training, the patient's facial image is captured by the digital camera. The patient needs to face the camera immovably according to the indicated contour on the screen. Then, the acquired image was tailored as 224×224 resolution by deep learning algorithms and serves as input data for video generation. The detailed manipulation can be reviewed in. The image would be displayed in the center of the screen. And the patient easily performed this procedure according to the instruction above the screen.

2.1.3. Speech synthesis engine

Speech synthesis, also known as TTS, is converting text into spoken speech. Machine learning techniques, particularly deep learning models, have revolutionized the field of speech synthesis, enabling the creation of natural-sounding and expressive synthetic voices.

Feature Extraction: After the dataset of paired text and audio samples is collected in the above step, the text input is transformed into a numerical representation, using word embedding. This numerical representation is fed into the speech synthesis model.

Model Training: During the training phase, a Speaker Verification (SV) algorithm is used for generating a fixed-length representation, called a speaker embedding, from audio files. The goal is to minimize the difference between the predicted speech and the ground truth audio from the dataset. Then, the TTS component is responsible for converting input text into speech. In an SV2TTS system, the speaker embedding obtained from the SV component is used as an additional input to the TTS system. Thus, the system can generate speech that sounds like the target speaker. It refers to for details of the SV2TTS system.

2.1.4. Video generation engine

Video generation with machine learning involves using deep learning models to create new, coherent and realistic video sequences. It is an extension of image generation and involves generating a sequence of video frames that flow seamlessly, capturing spatial and temporal dependencies to resemble real-world video footage.

Preprocessing: The images data is preprocessed to ensure a consistent format, resolution. The video frames are resized, normalized, and converted into a suitable format for the deep learning model.

Model Training: The selected deep learning model is trained on the preprocessed image dataset. During training, the model learns to capture the spatial information within each frame and the temporal dependencies between consecutive frames. The training process involves optimizing the model's parameters to minimize the difference between the generated video and the ground truth video from the dataset.

Sampling: After the model is trained, it can generate new videos. The model is given an initial frame or seed, and it autonomously generates subsequent frames, creating a coherent and continuous video sequence.

In our system, an integrated technology, called Pose-Controllable Audio-Visual System (PC-AVS), is used for video generation. The detailed information can be referred to. We have utilized the VoxCeleb2 and LRW datasets, which contain various facial expressions and head poses for model training. VoxCeleb2, with its over 1 million utterances, provides a rich source for understanding natural facial expressions across diverse conditions, including those with relevance to medical symptom expression. The LRW dataset, being cleaner and mostly containing frontal faces, allows for focused analysis of lip movements, which could be used for detecting abnormalities in speech patterns that may correlate with medical conditions.

2.1.5. Evaluation platform

The evaluation platform is also proposed to provide application software, as displayed in Fig. 2. The evaluation platform employs ARST algorithm for assessing the comparing between the patient's speech and the synthesized speech. And then the difference between the translated text and the standard text is used for scoring the patient's performance.

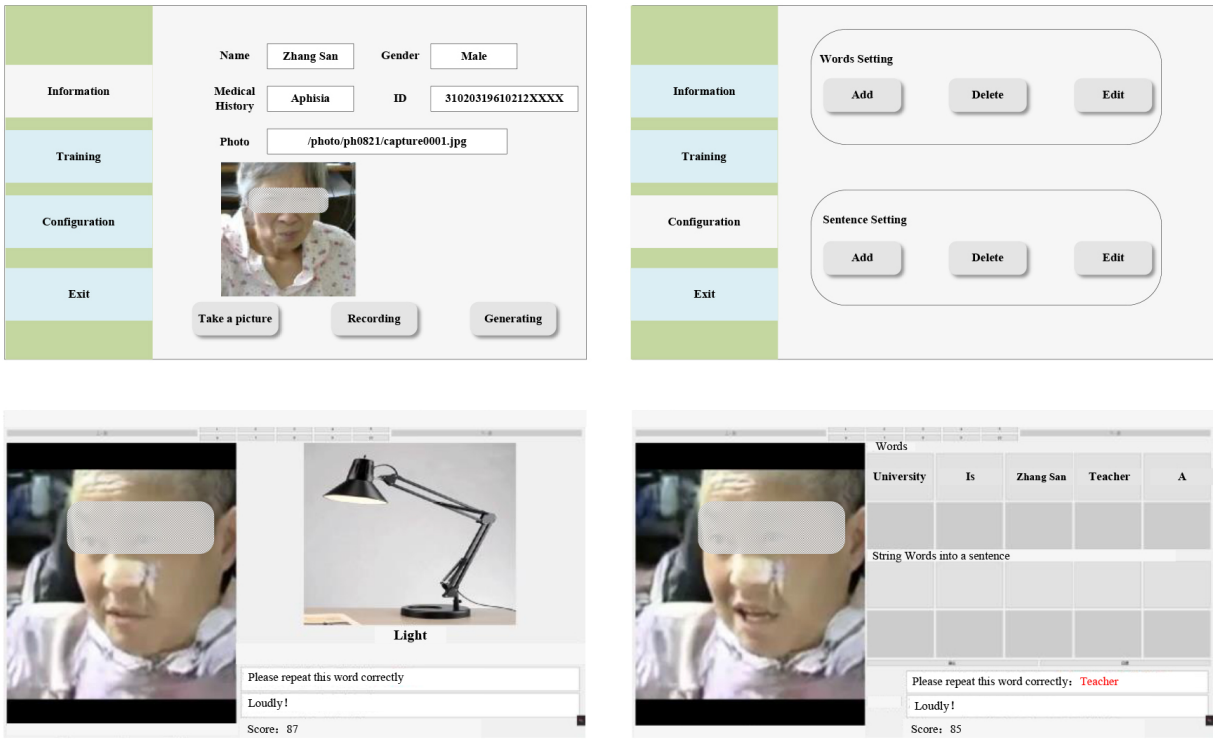


Fig. 2. Evaluation platform for rehabilitative training.

2.1.6. Training task module

Aphasia rehabilitation training involves two tasks and exercises to target specific language and communication deficits. The tasks used for aphasia rehabilitation training are tailored to each individual's type and severity of aphasia:

Word Exercises: Word exercises aim to improve word retrieval and naming abilities. The individual is presented with pictures of objects and corresponding words, the patient is asked to read them. In the same time, a face-mirror speaking video is displayed on the screen. Then, the patient needs to repeat this word following this video.

Stringing words together into sentences: The individual is given several words and asked to reconstruct them into a sentence. Simultaneously, a face-mirror video is displayed where the patients speak this sentence after 30 seconds. Then, the patient needs to repeat this sentence following the video. This exercise helps improve grammatical structure and sentence formation.

2.2. Task-oriented mirror therapy

The task-oriented mirror therapy experiments were performed in our study. Two types of experiments, naming exercises and sentence completion, were conducted for mirror therapy. 26 subjects (15 male and 11 female), ranged from 46 to 67, participated in our work. All of them gave written consent and were informed about the procedure in our study.

Figure 3 illustrated the experimental procedures. In the experiment of naming exercises, the patient was asked to name it according to the displayed picture in 10 seconds. Then, the mirror video was repeated 5 times. The subject needed to speak this vowel or word following this video in 5 seconds. Between these

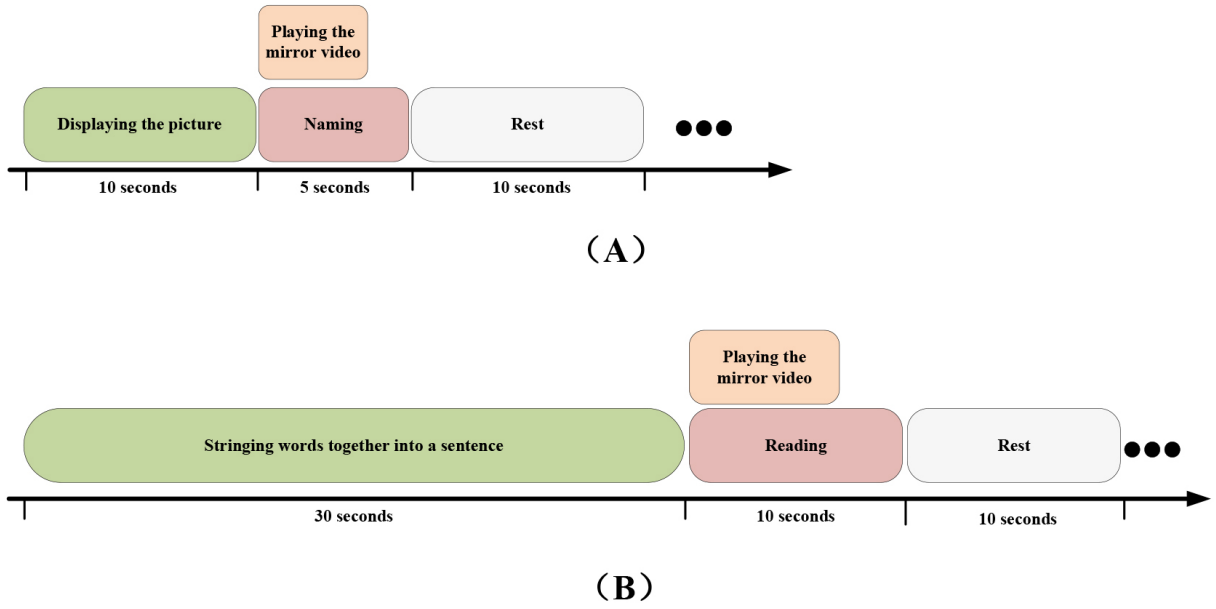


Fig. 3. (A) The experitmental procedures for naming exercises. (B) The experitmental procedures forStringing words together into sentences.

tests, the patient had a rest for 10 seconds. The patient needed to complete the test 20 times. In another experiment, 5–8 words were displayed out of order on the screen, the patient had to reconstruct them into a sentence and speak it in 30 seconds. After that, the mirror video was already repeated 5 times. And the patients needed to complete the repeating task in 10 seconds. In this experiment, the patient had to perform this task 10 times.

The software system was implemented in Python 3.7.6, running on the PC with 64-b Windows 7, 3.8-GHz Intel i7 processor, and 32 GB of RAM.

2.2.1. Video generating

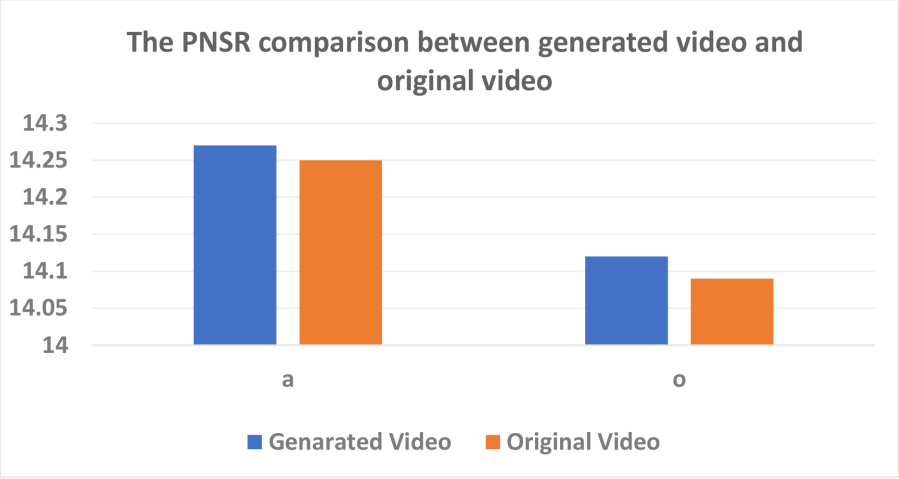
Before the experiment, the video generation engine would train a deep learning model by the subject's image. 2 indicators, structual similarity index (SSIM) and peak signal-to-noise ratio (PSNR), are used for evaluating the performance of this model. SSIM is a perception-based model that measures the similarity between two images. It considers the perceived changes in structural information, luminance, and contrast of the images. SSIM is a widely used method for measuring the similarity between two images or videos. It was developed to assess the quality of images or videos in a way that aligns more closely with human perception. Unlike traditional metrics that rely solely on pixel values, SSIM takes into account the structural information, luminance, and contrast that are important to human visual perception. The average SSIM score across all frames provides an overall measure of the similarity between the two videos. Higher SSIM scores indicate better similarity and quality. On the other hand, PSNR was a straightforward and objective method that quantifies the difference between the original and reconstructed frames in terms of pixel values and noise. The formula for PSNR is,

$$\text{PSNR} = 10 \cdot \log_{10} \frac{\text{MAX}}{\text{MSE}} \tag{1}$$

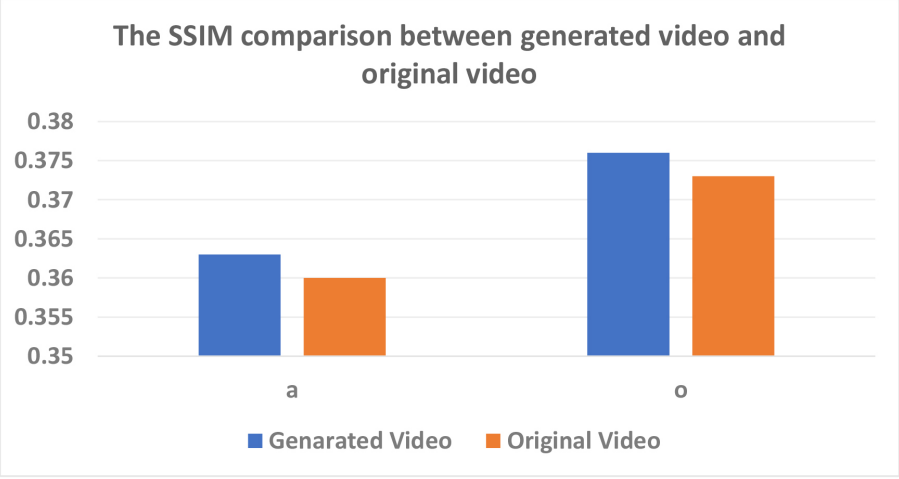
where MAX is the maximum possible pixel value of the image. MSE (Mean Squared Error) is the average squared difference between the original and compressed images. A higher PSNR value indicates lower pixel-wise differences between the original and generated frames, implying higher image quality.

Table 1
The Chinese vowels for naming exercies

Chinese vowel	a	o	e	i	u	v	ai	ei	ui	ao



(A)



(B)

Fig. 4. (A) The comparison between generated video and recording video for PNSR. (B) The comparison between generatied video and original video for SSIM.

We recorded the training videos for comparison with the generated video. Figure 4 showed these 2 indicators for all subjects. The performance of video generation engine was satisfied for generating real patients' face images. It was implied that the mirrored training was feasible for cognitive rehabilitation.

2.2.2. Chinese words and sentences

In the experiment, 33 words and 10 sentences were used for speech training. Tables 1–3 listed these Chinese vowels, words and sentences. Every sentence was splitted by several words and they would be displayed out os order in the second task.

Table 2
The Chinese words for naming exercies

Chinese word	图书馆	书桌	空调	黄色	床
Translation	Library	Desk	Air Conditioner	Yellow	Bed
Chinese word	窗户	绿色	冰箱	大学	游泳
Translation	Window	Green	Fridge	University	Swin

Table 3
The Chinese sentences for stringing words experiments

Chinese sentences	Translation
张三是一位大学老师	San Zhang is an university teacher.
我在图书馆学习	I am studying in the library.
这是我的眼睛	This is my eyes.
我想吃东西	I want to eat some food.
我需要一些帮助	I need help.
我需要休息	I need a rest.
我喜欢看电影	I like watching movies.
我要去医院	I have to go to the hospital.
我今天感觉很好	I feel good.
你喜欢什么颜色	What color do you like?

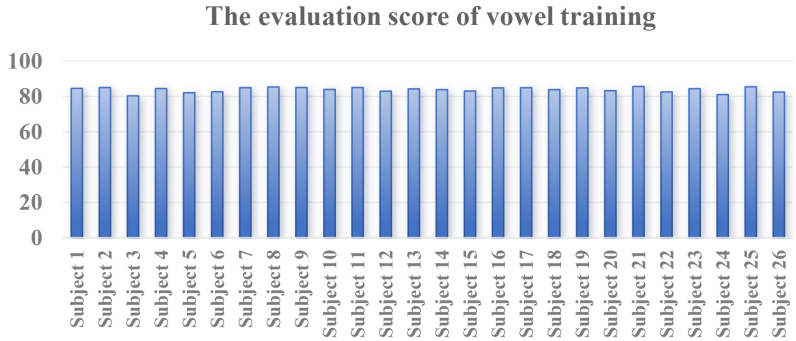


Fig. 5. The evaluation scores of vowel training by ARST algorithm.

3. Experimental results

3.1. Experiments for naming exercises

In the begining, the generated video was played as the instruction for the patient. Then, the subject had to imitate the former mirror expression for repeating the vowel or word. At last, the evaluating platform assessed the similarity between subjects' voices and critical voices generated by the speech synthesis engine.

Figures 5 and 6 presented the performance of similarity evaluation for all subjects. The results implied our system was helpful for patients to improve their pronunciations by these training.

3.2. Experiments for stringing words together into sentences

Firstly, the patient needed to realign these words into a complete sentence. Then the mirrored video was displayed for reading through. Then, the evaluating platform performed the score assessment as above.

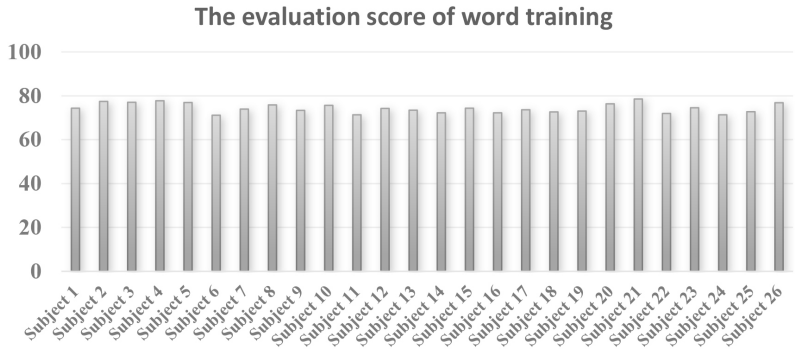


Fig. 6. The evaluation scores of word training by ARST algorithm.

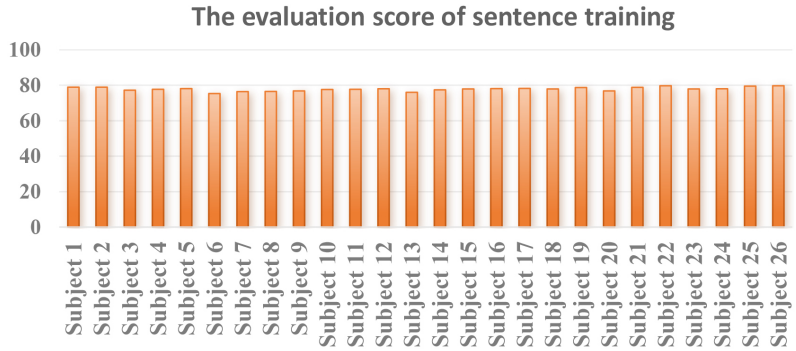


Fig. 7. The evaluation scores of sentence training by ARST algorithm.

Figure 7 reported the performance of sentence training for all subject. And the performance was implied that the mean correctness of sentence reading was lower than that of word reading. The results were in line with expectations.

The experimental section provides a clear demonstration of the implementation process and evaluation outcomes of the task-oriented mirror therapy. However, it is advisable for the authors to conduct a more comprehensive analysis of the experimental results. In further work, we would perform a prolonged experiment to adequately verify training effects over time.

Furthermore, in interpreting the experimental findings, it is recommended that the authors address potential limitations. These may include aspects such as sample size and participant-specific characteristics. Such insights would enhance the understanding of the results and provide valuable context for future research endeavors.

4. Conclusion

In this study, we proposed a mirror aphasia rehabilitation software system with artificial intelligence technologies. The system could generate real expressions and voices for rehabilitation training. Thus, the patients could imitate themself in the mirrored video. The user-friendly application system allows patients to carry out daily training tasks instructed by the therapists or the menu information. Our work demonstrated a promising intelligent mirror software system for reading-based aphasia rehabilitation.

Acknowledgments

The work was supported by the Project of Jiangxu Health Commission, China (No. Z2022012).

Conflict of interest

None to report.

References

[1] Nancy A, Eva K, Gonia J, Guylaine L, Christel B, Marc Y. How artificial intelligence (AI) is used in aphasia rehabilitation: A scoping review. Aphasiology. 2024; 38(2): 305-336.
[2] Stefaniak JD, Halai AD, Ralph MA. The neural and neurocomputational bases of recovery from post-stroke aphasia. Nature Reviews Neurology. 2020; 16(1): 43-55.
[3] Luca R, Leonardi S, Maresca G, Marilena FC, Latella D, Impellizzeri F, Maggio MG, Naro A, Calabrò RS. Virtual reality as a new tool for the rehabilitation of post-stroke patients with chronic aphasia: an exploratory study. Aphasiology. 2023; 37(2): 249-259.
[4] Schumacher R, Halai AD, Ralph MA. Assessing and mapping language, attention and executive multidimensional deficits in stroke aphasia. Brain. 2019; 142(10): 3202-3216.
[5] Anni P, Siponkoski S, Brownsett S, Copland S, Viljami D, Aleksi S. Hodological organization of spoken language production and singing in the human brain. Communications Biology. 2023; 6(1): 779-787.
[6] Bullier B, Cassoudesalle H, Villain M, Cogné M, Mollo C, De Gabory I, Dehail P, Joseph PA, Sibon I, Glize B. New factors that affect quality of life in patients with aphasia. Annals of Physical and Rehabilitation Medicine. 2020; 63(1): 33-37.
[7] Wilson SM, Entrup JI, Schneck SM, Onuscheck CF, Levy DF, Rahman M, Willey E, Casilio M, Yen M, Brito AC, Kam W, Davis LT, de Riesthal M, Kirshner HS. Recovery from aphasia in the first year after stroke. Brain. 2023; 146(3): 1021-1039.
[8] Gilmore N, Dwyer M, Kiran S. Benchmarks of Significant Change After Aphasia Rehabilitation. Archives of Physical Medicine and Rehabilitation. 2019; 100(6): 1131-1139.
[9] Brown SE, Scobbie L, Worrall Brady MC. A multinational online survey of the goal setting practice of rehabilitation staff with stroke survivors with aphasia. Aphasiology. 2023; 37(3): 479-503.
[10] Cruice M, Botting N, Marshall J, Boyle M, Hersh D, Pritchard M, Dipper I. UK speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. International Journal of Language & Communication Disorders. 2020; 417-442.
[11] Picano C, Quadrini A, Pisano F, Marangolo P. Adjunctive Approaches to Aphasia Rehabilitation: A Review on Efficacy and Safety. Brain Sciences. 2021; 41.
[12] Arheix-Parras S, Barrios SC, Python G, Cogné MI. Sibon I, Engelhardt M, Dehail P, Cassoudesalle H, Moucheboeuf G, Glize B. A systematic review of repetitive transcranial magnetic stimulation in aphasia rehabilitation: Leads for future studies. Neuroscience & Biobehavioral Reviews. 2021; 212-241.
[13] Fahmy EM, Elshebawy HM. Effect of High Frequency Transcranial Magnetic Stimulation on Recovery of Chronic Post-Stroke Aphasia. Journal of Stroke and Cerebrovascular Diseases. 2021; 105855.
[14] Low YA, Lindland K, Kirton A, Carlson HL, Harris AD, Goodyear BG, Monchi O, Hill MD, Dukelow SP. Repetitive transcranial magnetic stimulation (rTMS) combined with multi-modality aphasia therapy for chronic post-stroke non-fluent aphasia: A pilot randomized sham-controlled trial. Brain and Language. 2023; 236: 105216.
[15] Stahl B, Darkow R, von Podewils V, Meinzer M, Grittner U, Reinhold T, Grewe T, Breitenstein C, Flöel A. Transcranial Direct Current Stimulation to Enhance Training Effectiveness in Chronic Post-Stroke Aphasia: A Randomized Controlled Trial Protocol. Frontiers in Neurology. 2019.
[16] Georgiou AM, Phinikettos I, Giasafaki C, Kambanaros M. Can transcranial magnetic stimulation (TMS) facilitate language recovery in chronic global aphasia post-stroke? Evidence from a case study. Journal of Neurolinguistics. 2020; 55: 100907.
[17] Repetto C, Paolillo MP, Tuena C, Bellinzona F, Riva G. Innovative technology-based interventions in aphasia rehabilitation: a systematic review. Aphasiology. 2021; 35(12): 1623-1646.
[18] Zhang J, Zhong D, Xiao X, Yuan L, Li Y, Zheng Y, Li J, Liu T, Jin R. Effects of repetitive transcranial magnetic stimulation (rTMS) on aphasia in stroke patients: A systematic review and meta-analysis. Clinical Rehabilitation. 2021; 35(8): 1103-1116.

[19] Georgiou AM, Kambanaros M. The Effectiveness of Transcranial Magnetic Stimulation (TMS) Paradigms as Treatment Options for Recovery of Language Deficits in Chronic Poststroke Aphasia. Behavioural Neurology. 2022; 1-25.
[20] Devane N, Behn N, Marshal J, Ramachandran A, Wilson S, Hilari K. The use of virtual reality in the rehabilitation of aphasia: a systematic review. Disability and Rehabilitation. 2022; 1-20.
[21] Spaccavento S, Falcone R, Cellamare F, Picciola Glueckauf RL. Effects of computer-based therapy versus therapist-mediated therapy in stroke-related aphasia: Pilot non-inferiority study. Journal of Communication Disorders. 2021; 94: 106158.
[22] Mahmoud SS, Pallaud RF, Kumar A, Faisal S, Wang Y, Fang Q. A Comparative Investigation of Automatic Speech Recognition Platforms for Aphasia Assessment Batteries. Sensors. 2023; 23(2): 857.
[23] Agarwal S, Saxena V, Singal V, Aggarwal S. Deep Learning-Based Computer Aided Customization of Speech Therapy in Lecture Notes in Electrical Engineering. Applications of Artificial Intelligence and Machine Learning. 2021; 483-494.
[24] Cavanaugh R, Kravetz C, Jarold L, Quique Y, Turner R, Evans WS. Is There a Research-Practice Dosage Gap in Aphasia Rehabilitation. American Journal of Speech-Language Pathology. 2021; 30(5): 2115-2129.
[25] Hsieh YW, Lin HY, Zhu JD, Wu CY, Lin YP, Chen CC. Treatment Effects of Upper Limb Action Observation Therapy and Mirror Therapy on Rehabilitation Outcomes after Subacute Stroke: A Pilot Study. Behavioural Neurology. 2020; 1-9.
[26] Shan CL Li JA, Chen WL, Ye Q, Zhang SC, Xia Y, Yang X, Yuan TF. Aphasia rehabilitation based on mirror neuron theory: a randomized-block-design study of neuropsychology and functional magnetic resonance imaging. Neural Regeneration Research. 2019; 1004.
[27] Bai Z, Zhang J, Zhang Z, Shu T, Niu W. Comparison Between Movement-Based and Task-Based Mirror Therapies on Improving Upper Limb Functions in Patients With Stroke: A Pilot Randomized Controlled Trial. Frontiers in Neurology. 2019; 10.
[28] Anni P, Siponkoski S, Brownsett S, Copland S, Viljami D, Aleksi S. Hodological organization of spoken language production and singing in the human brain. Communications Biology. 2023; 6(1): 779-787.
[29] Krzysztof M, Szymon Z, Andrzej C. Comparison of the Ability of Neural Network Model and Humans to Detect a Cloned Voice. Electronics. 2023; 12(21): 4458-4472.