Sensors and Materials, Vol. 38, No. 4 (2026) 1829–1850
MYU Tokyo

S & M 4409

Virtual Speech-language Pathologist Establishment as Rehabilitation Assistive Tool for Mandarin-speaking Aphasia Patients

Hsiang-Yueh Lai,[1*] Tzu-Yao Chung,[1] Neng-Sheng Pai,[1] Chia-Chieh Hu,[1]
Ping-Tzan Huang,[2] Chien-Ming Li,[3] and Chia-Hung Lin[1**]

[1]Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung City 41170, Taiwan.
[2]Department of Biomechatronics Engineering, National Pingtung University of Science and Technology, Pingtung 91201, Taiwan.
[3]Tainan Municipal Hospital, Infectious Disease Division of Internal Medicine Department, Tainan City 701, Taiwan.

(Received May 6, 2025; accepted December 24, 2025)

Keywords: aphasia, treatment and language rehabilitation (TLR), Mandarin-speaking aphasia patient (MSAP), virtual speech-language pathologist (VSLP), mel frequency cepstral coefficient (MFCC), YOLOv9

Aphasia is a language and communication disorder, which is primarily caused by brain damage, leading to impairments in language proficiency and comprehension abilities. Aphasia patients may struggle with auditory comprehension, spoken expressions, reading comprehension, and writing skills at different disorder levels. The golden period for treatment and language rehabilitation (TLR) is within the first six months. If aphasia patients practice daily under the guidance of a speech-language pathologist (SLP), they can gradually achieve their rehabilitation goals. During Mandarin language rehabilitation, aphasic patients undergo treatment through one-on-one communication interactions with the SLP. Then, the SLP evaluates the patient's mandarin speech-language ability on a case-by-case basis using Mandarin Fluency Assessment Standards (MFAS). Therefore, rehabilitation plans can be personalized in accordance with the evaluation results with the aim of improving the patient's listening, speaking, and reading abilities. To overcome the concerns of limited manpower and time consumption in treating language disorders, we propose a rehabilitation assistive tool for Mandarin-speaking aphasia patients (MSAPs), enabling them to practice in their daily lives and home environment. By leveraging mixed reality (MR), a virtual SLP (VSLP) is implemented, and then, various interaction scenarios can be simulated to support self-rehabilitation and training. This rehabilitation assistive tool allows aphasia patients to express their daily needs while also automatically evaluating their speech-language abilities. This SLP assistive tool integrates a You Only Look Once, version 9 (YOLOv9)-based object detector and a mel frequency cepstral coefficient (MFCC)-based feature extractor with a one-dimensional (1D) convolutional neural network (CNN) to enable the automated assessment of speech-language proficiency. Hence,

*Corresponding author: e-mail: anne@ncut.edu.tw
**Corresponding author: e-mail: eechl53@gmail.com
https://doi.org/10.18494/SAM5724

ISSN 0914-4935
https://myukk.org/

through intensive, personalized, and repetitive self-training, MSAPs can effectively enhance rehabilitation outcomes.

1. Introduction

Aphasia is a common syndrome resulting from brain damage caused by stroke (cerebral infarction), traumatic brain injury (such as car accidents or impact), brain tumors, neurodegenerative diseases, or brain infections. Such damage can lead to impairments in the brain's abilities to process language messages, ultimately diminishing or completely disrupting an individual's communication abilities. Consequently, aphasia patients may experience difficulties in understanding or expressing language in their daily lives. The brain processes language through a complex physiological pathway: sound enters the ear → transmission to the brain's auditory cortex → language comprehension in Wernicke's area (posterior speech area) → transmission of language-related information through the arcuate fasciculus → speech production in Broca's area (anterior speech area) → control of tongue and mouth movements by the brain's motor cortex → speaking responses. The relevant functional areas of the human brain[1,2] are shown in Fig. 1. Any disruption in this neural pathway can result in language deficits, particularly when damage occurs in the above key regions, all of which are located in the left hemisphere. Stroke (cerebral infarction) and traumatic brain injury are the leading causes of aphasia. The severity levels and nature of language impairment vary depending on the specific speech areas, as outlined below.

- Wernicke's area damage: This area is responsible for language comprehension. Damage to the posterior speech area results in receptive aphasia (Wernicke's aphasia). While patients may have no issues with hearing or speech production, they experience significant difficulties in understanding language, leading to irrelevant responses, the creation of new or nonsensical words, and even a complete inability to comprehend spoken language.
- Broca's area damage: This type of aphasia is an expressive aphasia. Patients have normal language comprehension but experience significant difficulties in speech production, such as halting and fragmented speech, and speech composed of short phrases. Because of the difficulty in speaking, patients may become frustrated and reluctant to engage in conversations.

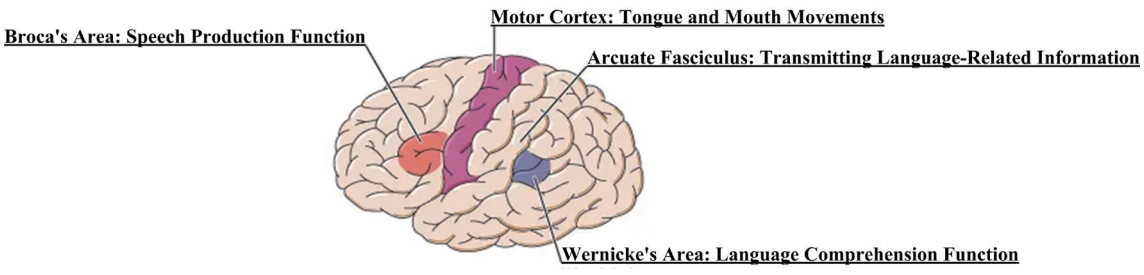


Fig. 1. (Color online) Functional areas of the human brain, including Wernicke's area, arcuate fasciculus, Broca's area, and motor cortex.

- Arcuate fasciculus damage: This type of aphasia is known as conduction aphasia. Patients have normal language comprehension but struggle with speech coherence and logical structure. They often have difficulty repeating words or phrases, despite understanding them. Since their comprehension remains intact, they are still able to accurately follow instructions and carry out requested actions.

Depending on the potential damage to the brain's language transmission pathways, aphasia may develop slowly over time and present with several common clinical symptoms, including: logopenic aphasia, semantic aphasia, and agrammatism.[3–5] Many patients with aphasia have some symptoms, as follows: (1) speak in short or incomplete sentences, (2) speak in sentences that do not make sense; (3) speak unrecognizable words; (4) substitute one word for another one or one sound for another word; (5) do not understand people's conversations or what they read. The goal of aphasia treatment is to establish effective communication between patients and others. This is achieved through comprehensive training in auditory comprehension, verbal expression, reading, and writing, gradually helping patients regain their speech-language abilities in daily life and self-expression.

According to statistics, Taiwan has approximately 65000 individuals diagnosed with aphasia, and as the aging population continues to increase (aged 65 and older), the prevalence of aphasia is expected to rise. In the golden period (first six months), treatment and language rehabilitation (TLR) is crucial and essential. However, there is a shortage of speech-language pathologists (SLPs), thus each SLP in Taiwan is responsible for approximately 100 aphasia patients, leading to an overwhelming demand for rehabilitation services. Consequently, many patients miss this crucial window for effective treatment and recovery. Among aphasia patients, stroke accounts for the highest proportion of long-term disabilities. Chronic aphasia is one of the most common sequelae in stroke patients, affecting approximately 10–18% of individuals. Language and communication disorders associated with chronic aphasia may significantly burden both patients and their families.[6–8] Hence, we propose a virtual SLP (VSLP) rehabilitation assistive tool for Mandarin-speaking aphasia patients (MSAPs) by combining the You Only Look Once, version 9 (YOLOv9)-based object detector (ODR),[9,10] mel frequency cepstral coefficient (MFCC)-based feature extractor,[11–13] one-dimensional (1D) convolutional neural network (CNN)-based classifier (1D-CNN classifier),[11–16] and mixed reality (MR) for MSAP rehabilitation. We intend to design a portable assistive tool that leverages various images or word cards to help MSAPs establish training scenarios and goals during the TLR stage. By integrating an image-text recognition and speech feedback assistive device, MSAPs can engage with MR glasses, facilitating a more intuitive and immersive language rehabilitation experience. Rehabilitators can select digital picture cards (for example, animal flashcards, such as elephant and giraffe); then, the proposed VSLP employs the YOLOv9 classifier to perform object detection (OD), after which the corresponding vocabulary associated with the recognized object is simultaneously presented on the MR display screen. The image–text database is instantly accessed to extract key vocabulary and generate language training phrases along with audio outputs, thereby facilitating speech-language training and enhancing communication with others. Additionally, rehabilitators can use a microphone to record their own voiceprint signals. Through a MFCC-based feature extractor and a 1D-CNN classifier, the proposed speech recognizer performs the

feature extraction and speech recognition tasks. Hence, it can automatically evaluate language ability and training effectiveness, and provide scoring feedback, achieving the function of assisting MSAPs in language learning and rehabilitation.

In this study, the YOLOv9-based classifier was trained and validated using the Microsoft Common Objects in Context (MS COCO) database for OD,[17] and the proposed 1D-CNN-based classifier was trained, tested, and validated using dialect-sentence speech corpora from Mandarin (MAN), American English (AE), Italian (IT), and Arabic (ArB) acoustic–phonetic continuous speech databases.[18–21] Using tenfold cross-validation, the experimental results demonstrated that the proposed VSLP, integrating a YOLOv9-based object detector and a speech recognizer composed of MFCC feature extractor + 1D-CNN classifier, could automatically identify both gender and regional accents from the extracted speech voiceprints by evaluating the metrics of precision (%), *recall* (%), *F*1 *score*, and *accuracy* (%),[13,22] Then, MSAPs' speech signals were converted into text, and Language Proficiency Scoring (*LPS*)[23] was used to evaluate their language ability through word-by-word *accuracy* analysis using the Google Speech Recognition (GSR) tool. Generating interactive scenario visuals through the MR interface as the so-called VSLP, as seen in Fig. 2, allowed MSAPs to respond and undergo speech-language proficiency assessment through real-time interactive audiovisual training scenarios and rehabilitation goals with the generation of virtual images, text, and spoken audio.

2. Materials and Methods

2.1 Related works

According to the American Speech-Language-Hearing Association (ASLHA), aphasia is a condition caused by damage to the brain's language centers, leading to varying degrees of acquired impairments in spoken expression, auditory comprehension, reading, and writing,[24,25] while speech clarity remains unaffected. According to the statistical data of Taiwan Speech-Language-Hearing Association (TSLHA), aphasia patients must seize the golden period for speech-language rehabilitation, that is, within six months. It is recommended that rehabilitation be facilitated through gestures, word cards, or picture cards, and written text; for example, picture and letter boards and speech-generating devices can help the individual express thoughts,

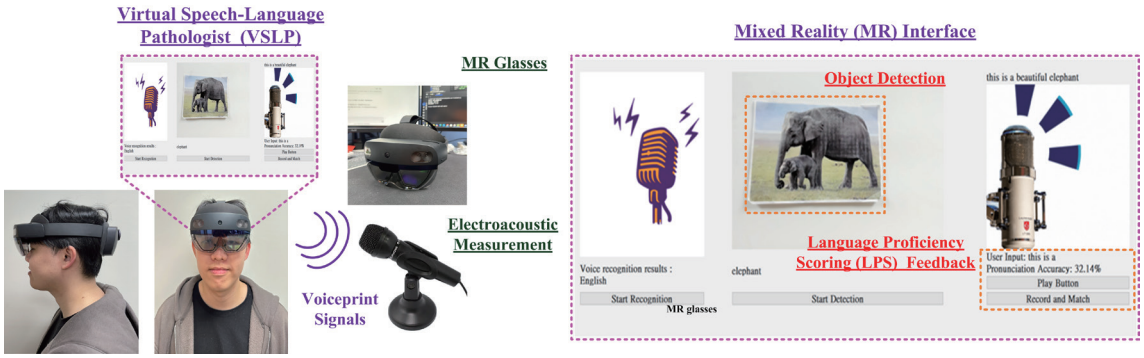


Fig. 2. (Color online) MR interface, the so-called VSLP.

wants/needs, feelings, and ideas.[26] The food and animal picture and letter boards in Fig. 3 offer commercially available sentence formation aids. Additionally, handheld communication devices, video-based tracking systems, touch-activated systems, and touch-based speech communication boards[27,28] can be effectively used to integrate rehabilitation into daily life and foster active engagement with patients' family and friends, significantly enhancing the recovery process. Aphasia classes can be divided into fluent aphasia and nonfluent aphasia. Symptoms of nonfluent aphasia include a low speech rate, increased effort in speech production, grammatical errors, short and fragmented sentences, disrupted rhythm, limited verbal output, and difficulty initiating conversations. During the first six months after onset, SLPs play a vital role in facilitating language rehabilitation. Aphasia intervention methods[26,29,30] include (1) melodic intonation therapy; (2) visual action therapy; (3) semantic feature analysis; (4) semantic-syntactic matching therapy; and (5) augmentative and alternative communication manners.

Throughout the rehabilitation processes, therapy progresses systematically, incorporating tasks of increasing complexity and programmed stimulation.

In this study, we design an AAC-based rehabilitation manner to improve speech-language disorders, engaging in one-on-one interactions with aphasia patients during each session. Using the patients' responses and scores, SLPs evaluate their progress and formulate subsequent treatment goals to gradually improve their speech-language abilities. However, the rehabilitation and training effectiveness is often impacted by various factors, including the limitations of coverage of Taiwan National Health Insurance (TNHI), time constraints for each session, and shortages of SLPs. Because of the aforementioned restrictions and session frequency limitations, providing intensive interaction or integrating therapy into daily life becomes increasingly challenging. Therefore, in this study, we propose a digital rehabilitation assistive tool (portable smart devices) for MSAPs, the so-called VSLP, to enable self-directed rehabilitation and training in a homecare setting at any time.

2.2 VSLP design and implementation

This assistive tool integrates a microphone, MR glasses, and a deep learning (DL)-based classifier to allow for personalized rehabilitation plans tailored to individual cases, incorporating high-density, repetitive practice with gradually increasing difficulty levels. Additionally, they

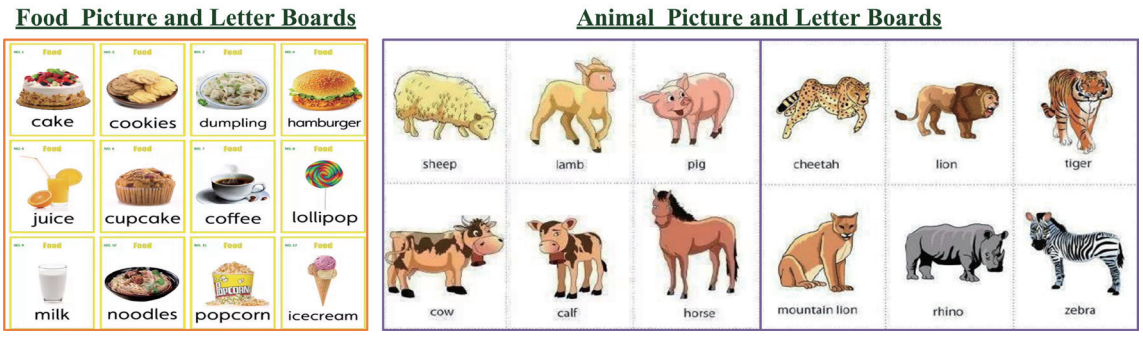


Fig. 3. (Color online) Food and animal picture and letter boards for speech-language rehabilitation.

can record practice results, enabling progress tracking and *LPS* responses. As seen in the flowchart in Fig. 4, we propose two main functions, as outlined below.

2.2.1 Function #1: MR to implement an AAC-based assistive tool for interactive speech-language training interface for MSAP

As seen in Fig. 4, the MR-based interface creates an interactive environment using a head-mounted transparent display (MR glasses). Through naked-eye vision, MSAPs can perceive the real-world environment while MR technology overlays digital content onto it, seamlessly integrating virtual and physical elements. This enables the MR platform to function as a "*VSLP assistant*" that presents vocabulary and sentence construction exercises within the MR environment. MSAPs can capture the picture and letter boards (e.g., animal/food picture and letter board in Fig. 4) using a camera integrated into the MR glasses and interact with the MR system through voice commands or gestures. In this study, we design a YOLOv9-based classifier for OD and object classification (OC)[9,10,31,32] within an image, identifying objects such as various foods or animals. Compared with the previous versions of YOLO models, YOLOv9 enhances the conventional backbone–neck–head configuration by integrating the feature pyramid network (FPN) and path aggregation network (PAN) to improve multiscale feature fusion and object localization. In addition, YOLOv9 also incorporates programmable gradient information (PGI) to facilitate better gradient flow and learning stability, thereby alleviating the vanishing gradient problem during deep network training, along with reversible convolutional (RevCol) operations, enabling invertible transformations that preserve input information and reduce memory requirements. Its model also employs the generalized efficient layer aggregation network (GELAN) to further optimize feature extraction efficiency and improve overall detection performance for objects at multiple scales. In the training stage, the stochastic gradient descent (SGD)-based optimizer[9,10,31,32] is applied to iteratively optimize the YOLOv9's network parameters. On the basis of specific objects depicted in the picture cards, the corresponding

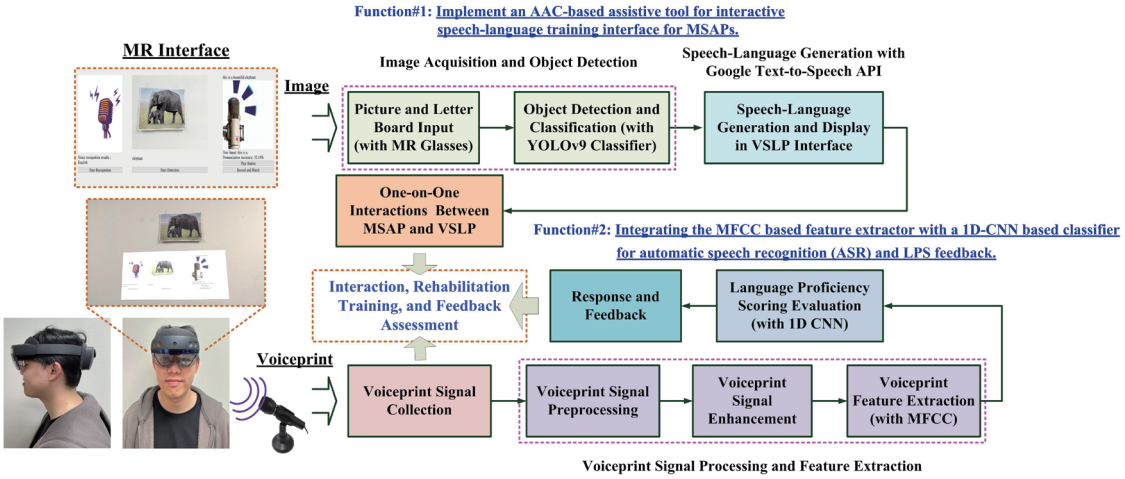


Fig. 4. (Color online) Flowchart of speech-language ability assessment for MSAPs.

short phrase constructions are generated, as shown in Table 1, which can automatically convert text into natural speech (human-like speech) using Google Text-to-Speech (GoogleTS) Application Programming Interface (API) + Python Playsound Module. Function#1 allows MSAPs to easily engage in an MR interactive environment with the VSLP progressing gradually through speech-generating exercises in accordance with predefined training goals.

2.2.2 Function #2: Integrating the MFCC-based feature extractor with a 1D-CNN-based classifier for automatic speech recognition (ASR) and *LPS* feedback

MSAPs use a high-quality USB microphone (specifications: directional microphone; sampling rate: 48 kHz; bit depth: 16 bits; frequency response range: 20–20000 Hz; power supply: USB 5 VDC) for audio recording and storage using the sounddevice + scipy.io.wavfile Python utility program. Within the MR environment, they participate in repetitive oral reading exercises and synchronized speech training using the proposed VSLP, facilitating personalized speech-language skill training and assessment. The MFCC-based feature extractor with a 1D-CNN-based classifier[22,33,34] is employed to perform the ASR for identifying the "Gender" and "Region", including MAN, AE, IT, Spanish (SP), ArB, and Japanese (JAP). Depending on the "Gender" and "Region," the VSLP leverages a speech database of short phrase constructions to generate reference spoken language signals, and then compares the reference voiceprint signal with the MSAP's voiceprint signal to facilitate ASR and *LPS* feedback. This rehabilitation assistive tool supports language training, communication practice, and the development of comprehension skills through short sentence exercises. The *LPS* assessment is conducted using the mean opinion score (MOS) established by the SLP, which is used to assess the overall perceived quality of a speech and an audio sample. The MOS is based on human ratings, where evaluators assess perceptual quality on a scale from 1 to 5, as shown for five speech quality categories in Table 2.[34,35] Therefore, the VSLP implementation can achieve the language training, communication training, and comprehension skill development goals. Using MR and Playsound Module, the VSLP can display picture and letter boards, short phrase constructions, and spoken audio signals. The enrolled MSAP first listens to the short phrase read by Playsound Module, then practices reading it aloud or repeating it multiple times. The MSAP's spoken audio signals are used to assess training effectiveness and provide feedback, as seen in Fig. 4. The MSAP can engage in interactive communication training with the VSLP in simulated real-life scenarios, and gradually enhance vocabulary retrieval, sentence structure expression, and listening comprehension skills.

2.3 MFCC feature extractor design

In clinical speech-language practice, the Concise Chinese Aphasia Test (CCAT) is commonly used to provide a traditional quantitative manner of assessing the rehabilitation effectiveness.[36] The CCAT includes 12 evaluation scores to comprehensively assess different aspects of language abilities in individuals with aphasia, as seen in the 12 scores and their corresponding responses in Table 3, including simple responses, oral narration, picture-object matching, auditory

Table 1
Different animal categories and their corresponding short phrase constructions.

Board no.	Animal categories	Short phrase constructions (MAN/AE)
1	Bird (鳥)	This is a beautiful “bird”. (這是一隻美麗的 ” 鳥 ”)
2	Cat (貓)	This is a tabby “cat”. (這是一隻花 ” 貓 ”)
3	Dog (狗)	A “dog” is chasing a car. (一隻 ” 狗 ” 在追車子)
4	Horse (馬)	A “horse” is on the ranch. (一隻 ” 馬 ” 在牧場)
5	Sheep (羊)	A “sheep” is strolling across the grassland. (“ 羊 ” 正在草原上漫步)
6	Cow (牛)	A “cow” is grazing. (一隻 ” 牛 ” 正在吃草)
7	Elephant (大象)	This is a strong “elephant”. (這是一隻強壯的 ” 大象 ”)
8	Bear (熊)	This is a black “bear”. (這是一隻黑色的 ” 熊 ”)
9	Zebra (斑馬)	A “zebra” is a striped animal. (“ 斑馬 ” 是有條紋的的動物)
10	Giraffe (長頸鹿)	A “giraffe” has a very long neck. (” 長頸鹿 ” 脖子很長)

Table 2
Five speech quality categories and their MOS and degradation rating scales.

Rating	Speech quality category	Degradation	*LPS*
5	Excellent	Imperceptible	100–80%
4	Good	Just perceptible but not annoying	80–60%
3	Fair	Perceptible, slightly annoying	60–40%
2	Poor	Annoying	40–20%
1	Unsatisfactory	Very annoying	0–20%

Table 3
CCAT scoring system.

Score	Assessment focus	Evaluation criteria
01	Simple responses	Ability to provide basic answers to simple questions
02	Oral narration	Ability to describe pictures, events, or situations fluently
03	Picture–object matching	Accuracy in linking spoken words to the correct images or objects
04	Auditory comprehension	Understanding of spoken instructions or conversations
05	Word expression	Ability to retrieve and articulate appropriate words
06	Reading comprehension	Understanding of written text, phrases, or sentences
07	Sentence repetition	Ability to accurately repeat sentences of varying lengths and complexity
08	Writing ability	Capable of writing words, phrases, or sentences
09	Functional communication	Effectiveness of real-world communication skills
10	Grammar and syntax	Correct use of grammar, sentence structure, and word order
11	Fluency and pronunciation	Speech fluency, articulation clarity, and pronunciation accuracy
12	Overall communication ability	General assessment of language proficiency and effectiveness in everyday interactions

comprehension, word expression, reading comprehension, sentence repetition, and others. The test uses two parallel forms (Form #A and Form #B) administered alternately to evaluate the MSAP’s language recovery, progress, and treatment outcomes. However, this traditional one-on-one communication-based rehabilitation is time-intensive and human-resource-demanding. Therefore, the ASR, including feature extraction, feature quantification, and voiceprint recognition, can significantly enhance the efficiency of voiceprint processing. The ASR relies on key feature parameters as the foundation for voiceprint recognition. The MFCC-based feature extractor[13] serves to obtain appropriate mel-scale feature parameters for voiceprint recognition,

as seen in the flowchart of MFCC feature extraction in Fig. 5. The discrete cosine transform (DCT) is the ability to simplify frequency domain transformations using only real-number operations,[13] thereby reducing computational complexity. After the MFCC transformation, the unique biological characteristics of voiceprint signals can be visualized as colorful feature patterns, as seen in the gender and different regional feature patterns (AE, IT, and SP) in Fig. 5. For example, with 60-frame processing where each frame contains 13 feature parameters, possible DC components or high-frequency elements that do not contribute to classification are eliminated, thus a 60 × 13 feature pattern can be reduced to a 13 × 13 size, and the visualized feature pattern can then be normalized.

In voiceprint feature extraction, the MFCC is used to extract key parameters from speech signals. The process workflow includes "Signal Filtering + Endpoint Detection (ED) + Signal Pre-emphasis + MFCC Processing", which is used to extract frequency-related features, such as the mel–scale frequency and amplitude, for differentiating gender and regional speech dialects. Table 4 shows the results of speech signal preprocessing and feature extraction.

2.4 1D-CNN-based classifier design

The DL-based classifier can be trained to automatically perform the feature extraction, enhancement, and pattern recognition tasks.[11–16] By applying multiple layers of convolutional operations with sliding convolutional windows at each layer, a weighted combination of convolutional kernels with varying weights is generated, which enhances the depth and breadth of feature patterns, expanding feature dimensions, increasing nonlinearity, and capturing more complex patterns. This enhances the classifier's ability to recognize complex feature patterns. Then, a pooling process is used to select key feature parameters by maximum pooling (Max-pooling), and the number of feature parameters is reduced to one-fourth that of the original

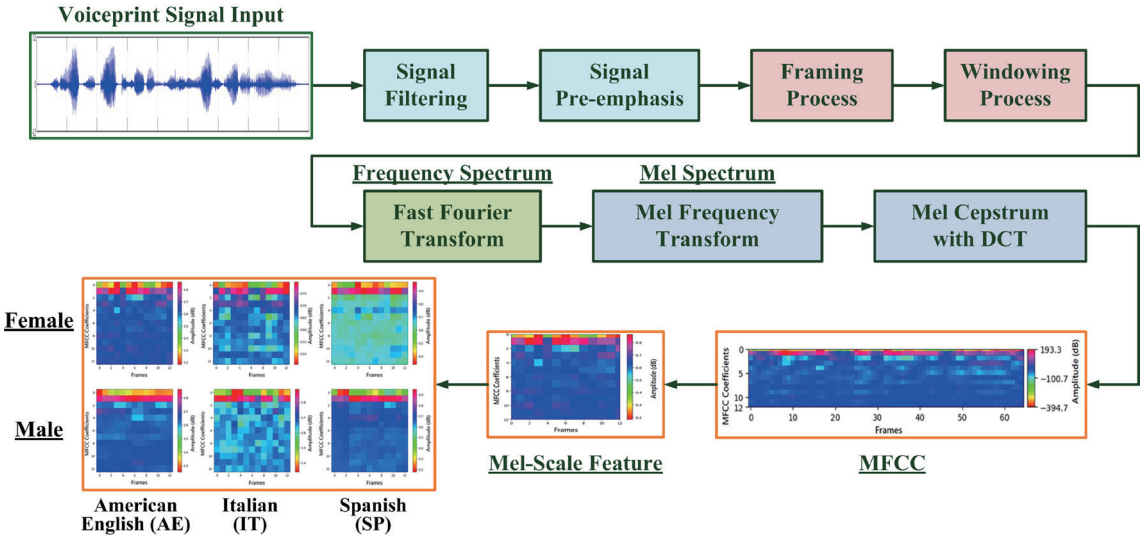
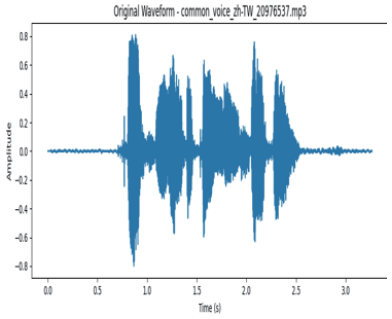
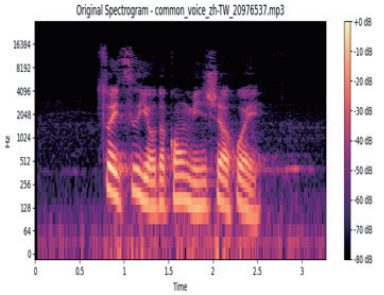
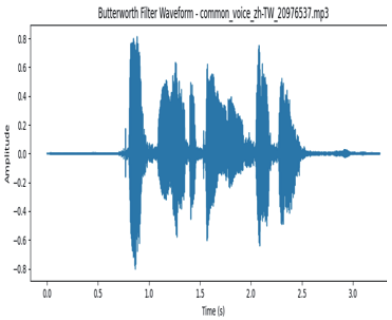
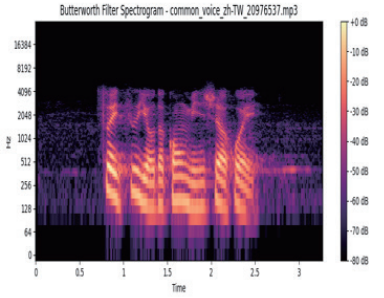
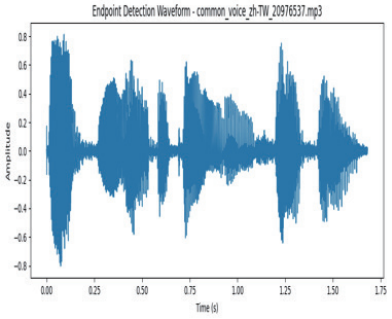
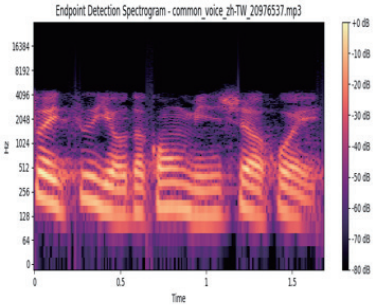
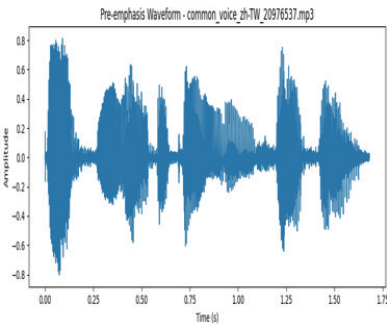
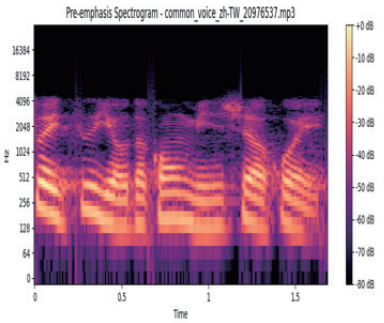


Fig. 5. (Color online) Flowchart of MFCC transformation.

Table 4
(Color online) Results of speech signal preprocessing and MFCC feature extraction.

Method	Processed speech signal	MFCC feature extraction
MFCC		
Signal filtering + MFCC		
Signal filtering + ED + MFCC		
Signal filtering + ED + signal pre-emphasis + MFCC		

feature pattern while preserving the essential characteristics of the original feature pattern. Therefore, CNN can directly extract key features and recognize different patterns from MFCC feature maps. In the classification layer (fully connected layer), the dense network is trained using optimization learning algorithms, such as the backpropagation or adaptive moment estimation (ADAM) algorithm.[13,22]

Traditional 2D-CNNs have some drawbacks and limitations,[37–39] including the need to determine the optimal number of convolutional-pooling (Conv-Pool) layers, the computational complexity of the training process, and the requirement for large amounts of training parameters. Additionally, 2D-CNNs rely on graphics processing units (GPUs) to accelerate the computation process. To address these limitations, 1D-CNN offers an alternative solution, particularly for processing and classifying 1D digital signals, such as electrocardiogram (ECG) signals, blood pressure waveforms, and vibration signals.[37,38] Compared with 2D-CNNs, 1D-CNNs have lower computational requirements and complexity, making them easier to implement on computers or embedded systems. Hence, in this study, we construct a 1D-CNN-based classifier for identifying gender and regional speech dialects, consisting of two 1D Conv-Pool layers, one flattening layer, and three dense networks, as seen in Fig. 6. The 1D-CNN-based classifier is developed using TensorFlow (Google Brain Team, 2015) with the high-level Keras API (a free and open-source software library for classifier modeling) to implement in Python. The dense (fully connected) network comprises an input layer, hidden layers, and an output layer, with the Gaussian error linear unit (GeLU) activation function selected for the hidden layers. The ADAM-based optimizer[13,22,40] is used to iteratively adjust the classifier's weight parameters through iteration computations to minimize the loss function (LF). The LF for multiclass classification is defined as

$$L = -\frac{1}{K}\sum_{j=1}^{m}\sum_{k=1}^{K} t_{j,k}\log_2(y_{j,k}) + (1-t_{j,k})\log_2(1-y_{j,k}), \tag{1}$$

```
scores = model.evaluate(test_value_4D, test_label_OneHot)
print('acc=', scores[1], '\nloss=', scores[0])

f = open(os.path.join(now_path, 'train_history', 'train_history.txt'), 'a+') # a+

content = f.read()
f.seek(0, 0)

f.write('\n' + '--------------------------------' + '\n')
f.write('\n' + '*****' + Tr_ver + '*****' + '\n')
f.write(now_str + '\n')
f.write('acc=' + '\n' + str(scores[1]) + '\n')
f.write('loss=' + '\n' + str(scores[0]))

f.write('\n' + content)
f.close()
```

```
798/798 [==============================] - 0s 229us/sample - loss: 0.1088 - acc: 0.9718
acc= 0.9718045
loss= 0.10876718034882493
```

```
# 使用pd.crosstab建立混淆矩陣，輸入下列參數:
prediction = model.predict_classes(test_value_4D)
pd.crosstab(test_label, # 測試資料數字影像的其實值
            prediction, # 測試資料數字影像的預測結果
            rownames=['label'], # 設定行的名稱是label
            colnames=['predict']) # 設定列的名稱是predict
```

Layer (type)	Output Shape	Param #
conv1d_14 (Conv1D)	(None, 169, 32)	128
max_pooling1d_14 (MaxPooling	(None, 84, 32)	0
conv1d_15 (Conv1D)	(None, 84, 64)	6208
max_pooling1d_15 (MaxPooling	(None, 42, 64)	0
flatten_7 (Flatten)	(None, 2688)	0
dense_21 (Dense)	(None, 128)	344192
dropout_14 (Dropout)	(None, 128)	0
dense_22 (Dense)	(None, 32)	4128
dropout_15 (Dropout)	(None, 32)	0
dense_23 (Dense)	(None, 3)	99

Total params: 354,755
Trainable params: 354,755
Non-trainable params: 0

Fig. 6. (Color online) TensorFlow development and design interface for 1D-CNN implementation.

where $t_{j,k}$ represents the target value (actual value); $y_{j,k}$ represents the predicted value output by the classifier; j = 1, 2, 3, …, m and k = 1, 2, 3, …, K, m is the number of classes, and K is the number of training datasets. We collect four speech corpora, MAN, Texas Instruments/ Massachusetts Institute of Technology (TIMIT), Corpora e Lessici dell' Italiano Parlato e Scritto (CLIPS), and Arabic Acoustic-Phonetic Continuous Speech Corpus,[18–21] to develop an ASR classifier model for four language systems—MAN, AE, IT, and ArB. The classifier is designed to automatically recognize the speaker's gender and regional accents from the given different dialect voiceprint signals.

2.5 Classifier performance and speech-language proficiency evaluation

2.5.1 Classifier performance evaluation

We employ tenfold cross-validation to evaluate the effectiveness of the proposed method, implementing MFCC + 1D-CNN + ADAM for ASR. The classifier's performance is evaluated using the *F*1 *score* metric, which serves as the harmonic mean of *Precision* (%) and *Recall* (%), offering a balanced measure of both aspects.[13,22] A higher *F*1 *score*, approaching 1, indicates the superior predictive performance of the classifier. The mathematical formulas are

$$Sensitivity\,(\%) = Recall\,(\%) = \left(\frac{TP}{TP+FN}\right)\times 100\%, \tag{2}$$

$$Specificity\left(\%\right) = \left(\frac{TN}{TN+FP}\right)\times 100\%, \tag{3}$$

$$Precision\left(\%\right) = \left(\frac{TP}{TP+FP}\right)\times 100\%, \tag{4}$$

$$F1\ score = \frac{2TP}{2TP+FP+FN}, \tag{5}$$

where *TP* represents true positive; *FP* represents false positive; *TN* represents true negative, and *FN* represents false negative. The classifier generates a confusion matrix composed of these four indicators (*TP*, *FP*, *TN*, and *FN*). This confusion matrix is then used to calculate the metrics in Eqs. (2) to (5) to evaluate the classifier's performance.

2.5.2 Speech-language proficiency evaluation

In this study, on the basis of predefined training scenarios, the specific sentence samples were set for each training goal and context. The process is divided into three stages.

- Step #1) The VSLP automatically identifies the target objects on the picture and letter boards with the YOLOv9-based classifier and then converts their corresponding text into natural (human-like) speech (MAN default) using GoogleTS + Python Playsound, which can generate expressive voices capable of conveying a wide range of emotions and intonations. YOLOv9 is the well-known model for OD and OC within an input image.[31,32] In addition, the text-to-speech tool enables adjustments to speech speed and pitch, as well as modifications to the volume of the generated voice speech.
- Step #2) The 1D-CNN-based classifier is used to automatically recognize the gender and region accents. Then, on the basis of the MSAP's gender and native language, the text-to-speech tool generates accurate speech signal samples from the speech library (MAN, AE, IT, and ArB). The MSAP visually reviews and reads aloud short phrase combinations while simultaneously recording and capturing speech audio.
- Step #3) Using the GSR tool, the MSAP's speech signal undergoes speech-to-text conversion and is analyzed for word-by-word *accuracy* by comparing it to the generated natural speech signal based on the MSAP's native language. The *LPS* can be mathematically expressed as

$$LPS\,(\%) = \frac{Match}{Length\ of\ Text} \times 100\%\,, \tag{6}$$

where *Match* represents the number of correctly matched words in the transcription; *Length of Text* refers to the total number of words in the reference text. Equation (6) provides a quantitative measure for speech-language proficiency evaluation. The training process of speech-language proficiency is shown in Fig. 7.

2.6 MR interactive interface design

MR can enable one-to-one or one-to-many modes of instructional guidance. Through the MR glasses projecting virtual objects, the participant can view virtual content in the real world, as seen in Fig. 2, enabling the participant to intuitively interact with assistive devices. This interaction follows predetermined training goals, allowing the participant to sequentially

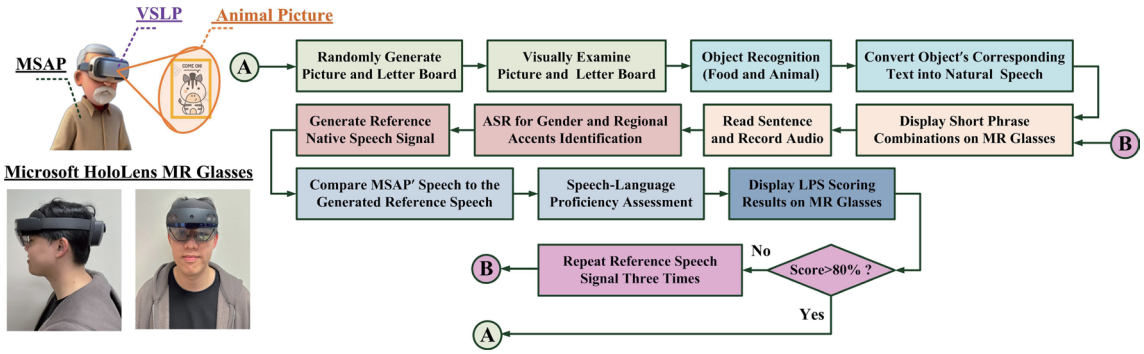


Fig. 7. (Color online) Training process of speech-language proficiency.

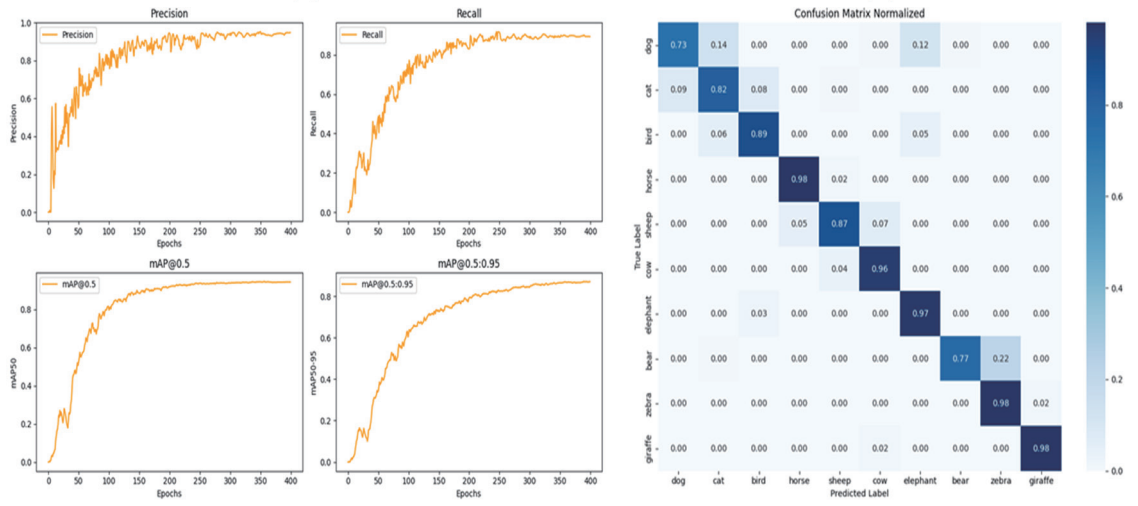
perform each rehabilitation task. This approach significantly reduces the time and manpower costs of instructors. We develop an MR-based SLP interactive interface, the VSLP, using the Microsoft HoloLens [operating system: Windows Mixed Reality, CPU: Intel Atom x5-Z8100 (1 GHz), memory: 2 GB RAM, sound: spatial sound technology, camera: 2.4 MP], as shown in Fig. 8(a). Microsoft HoloLens is a standalone holographic device platform that integrates spatial positioning, object recognition, speech recognition, and audiovisual effects within the real world. Additionally, we also incorporate the YOLOv9 model[9,10,31,32] to achieve the OD and OB tasks, for example, using animal picture and letter boards, as seen in Fig. 8(b). Ten animal categories are selected from the MS COCO database as the target objects for training and testing of the YOLOv9-based classifier. The training datasets (2542) and testing datasets (1570) are shown in Table 5, and the validation history curves are shown in Fig. 9(a), where the *mAP*_50 and *mAP*_50-95 progress in YOLOv9 across 400 iterations can be seen for the evaluated object localization and classification tasks. The YOLOv9 model's recognition results for the ten animal categories are shown in the normalized confusion matrix, as seen in Fig. 9(b). The YOLOv9 model can accurately recognize specific objects such as cows, elephants, bears, and zebras within the picture board, as seen in Fig. 8(b). By integrating speech and visual objects, the proposed VSLP can form an interactive interface for MSAPs, enabling personalized

(a) (b)

Fig. 8. (Color online) Microsoft HoloLens MR glasses and automatic object recognition function (with YOLOv9-based classifier). (a) Microsoft HoloLens operating system and (b) animal picture and letter boards.

Table 5
Ten animal-category datasets for training and testing datasets.

Board no.	Animal category	Training dataset	Testing dataset
1	Bird (鳥)	250	156
2	Cat (貓)	252	162
3	Dog (狗)	268	160
4	Horse (馬)	260	158
5	Sheep (羊)	253	160
6	Cow (牛)	250	160
7	Elephant (大象)	250	155
8	Bear (熊)	261	160
9	Zebra (斑馬)	248	143
10	Giraffe (長頸鹿)	250	156



(a) (b)

Fig. 9. (Color online) (a) Validation history curves for evaluating the object localization and classification tasks. (b) The YOLOv9 classifier outputs a normalized confusion matrix for ten animal categories

rehabilitation training. The VSLP includes (1) the establishment of rehabilitation training goals; (2) personalized text and speech prompts and speech-language ability assessment; (3) personalized training scenario design, allowing the MSAP to engage in real-time interactive audiovisual scenarios (generating virtual objects, text, and spoken audio) and enabling them to respond and undergo language proficiency assessment.

3. Experimental Results

3.1 Speech corpus database collection

Natural language (NL) is a tool for communication and interaction in human social activities, such as transmission, expression, description, cultural preservation (inheritance), and information exchange. The same language may develop regional dialects over time owing to differences in countries and geographical areas. Even within the same region, variations in language can arise depending on gender, age groups, and social classes, incorporating honorifics or specialized vocabulary for communication. For example, the TIMIT Speech Database (The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, USA)[13,19] had collected a total of 6300 dialect sentences from eight major dialect regions in the United States (recording speech signals with 16-bit and 16 kHz sampling frequency). This speech database captures the diversity of American dialects and can be used for gender differentiation and regional dialect recognition, as well as for validating and evaluating the performance of the proposed ASR method. Used for distinguishing gender and different age groups, the average fundamental frequency (F_0) was as follows: adult females 216.78 ± 1.83 Hz; adult males 149.56 ± 4.27 Hz; children 233.12 ± 1.75 Hz (boys 232.17 ± 2.03 Hz; girls 234.07 ± 1.48 Hz).[13,19]

The CLIPS Corpus[20] was collected and recorded by the Italian Ministry of Education, University and Research (Ministero dell'Università e della Ricerca, MIUR). This corpus was recorded at 8-bit and 8 kHz sampling frequency and included linguistic diversity from 15 cities in Italy, such as Rome, Florence, Milan, Venice, and Parma. The dataset covers daily conversations, television broadcasts, telephone communications, speeches, and other linguistic sources. The participants ranged in age from 23 to 50 years old, with a total recording time of 100 h, highlighting the diversity of Italian language usage across different regions and contexts. Other notable corpora included the Mandarin Chinese Corpus (National Academy for Educational Research, 2024)[18,41] and Arabic Corpus (ArabiCorpus: The Quranic Arabic Corpus _ Word by Word Grammar, Syntax, and Morphology of the Holy Quran).[21] Hence, we collected four speech corpora, including Mandarin Chinese (MAN), TIMIT (AE), CLIPS (IT), and ArB, and then collected speech signals from each corpus for digital signal preprocessing (noise filtering), ED, pre-emphasis, MFCC feature extraction, and feasibility evaluation for training and validating the ASR classifier.

3.2 Training and validation of ASR classifier

For the MAN, AE, IT, and ArB speech corpora, the mel-scale feature parameters were extracted from the speech signals. These time-domain feature parameters were then converted into visual representations, shown as the colored feature patterns in Fig. 5. These colored visual feature patterns were used to observe the differences in speech signal characteristics across different genders and nationalities. By analyzing these differences, adjustments could be made to digital signal preprocessing and the selection of key parameter ranges. Specifically, speech signals were selected from the four corpora, as shown in Table 6, including the MAN dataset[42] with 250 samples (150 training datasets, 100 testing datasets); AE dataset with 250 samples (150 training datasets, 100 testing datasets); IT dataset with 263 samples (163 training datasets, 100 testing datasets); and ArB dataset with 240 samples (138 training datasets, 100 testing datasets). The overall datasets were divided into a training dataset and a testing dataset, and then colored visual feature patterns were extracted from the four aforementioned speech corpora. The signal processing included signal filtering + ED + signal pre-emphasis + MFCC, as shown in Table 4. The colored visual feature patterns were applied to train, test, and validate the ASR classifier, which was used for automatic nationality and gender recognition. We utilized a Tablet PC (Intel® Xeon® CPU E5-2620 v4, 2.1 GHz, and 64 GB of RAM) as the development platform to implement the MFCC + 1D-CNN algorithm, as shown in Fig. 6, and a GPU (NVIDIA® GeForce® RTX™ 2080 Ti, 1755 MHz, 11 GB GDDR6) was used to accelerate the digital signal processing and pattern recognition, which was developed using the TensorFlow platform to

Table 6
Four speech datasets for training and testing datasets.

Dataset	MAN	AE	IT	ArB
Training dataset	150	150	163	138
Testing dataset	100	150	100	102
Total	250	250	263	240

design the 1D-CNN based classifier, additionally enabling the generation of an automatic inference engine system, seen as the human–machine interface in Fig. 6.

For ASR implementation, we employed a 1D-CNN to construct and train a classifier model capable of recognizing both gender and nationality from the MSAP's voiceprint signals. As shown in Fig. 6, the 1D-CNN-based classifier consisted of multiple layers, including one MFCC feature extraction layer, two 1D Conv-Pool layers, one flattening layer, and three dense networks. Between two dense layers, the dropout layer randomly deactivates 10% of neurons to reduce model complexity and mitigate the overfitting problem during the training stage. The ADAM algorithm[13,22] was employed to optimize the classifier's parameters by iteratively adjusting the network parameters to minimize the LF. As shown in Fig. 10, the training and validation history curves indicated that the 1D-CNN classifier could automatically recognize gender (0: Female, 1: Male) and nationality (0: MAN, 1: AE, 2: IT, 3: ArB). The maximum number of iterations was set to 350 epochs. As the number of iterations increases, the classifier's classification *accuracy* improves, while the LF gradually decreases and eventually converges to a specific condition, as shown in Figs. 10(a) and 10(b), respectively, where the blue line represents the training history curves and the orange line represents the validation history curves. The 1D-CNN classifier reached *accuracy* saturation after approximately 150 epochs. At the 350th epoch, it achieved an *accuracy* of 99.50%, with the LF converging to 0.0017, as shown in Fig. 10.

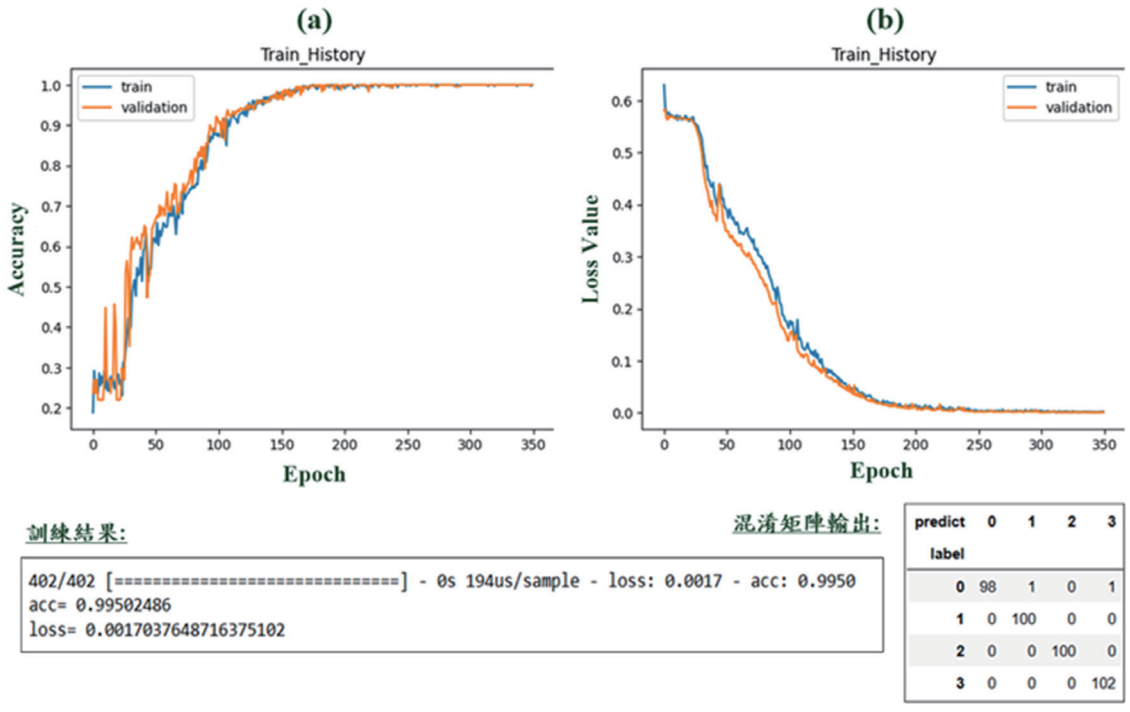


Fig. 10. (Color online) Accuracy saturation history curves and convergence history curves. (a) Accuracy versus training epoch history curves; (b) LF value versus training epoch history curves.

In this study, we employed tenfold cross-validation to evaluate the effectiveness of the proposed method, implementing MFCC + 1D-CNN classifier + ADAM optimization for ASR. In each validation fold, the classifier was trained and tested using the training and testing datasets, as shown in Table 7. Once the classifier model was trained, it generated a confusion matrix, as shown in Fig. 10. The confusion matrix outputs four key indicators: *TP*, *TN*, *FP*, and *FN*. On the basis of these index values, the performance metrics such as *Precision* (%), *Recall* (%), *F*1 *score*, and *Accuracy* (%) were computed to evaluate the feasibility of the 1D-CNN-based classifier. As shown in Tables 7 and 8, the experimental results from tenfold cross-validation yield an average *accuracy* of 99.48% and an average LF of 0.0102, confirming the classifier's capability to automatically recognize both the gender and nationality of incoming speech signals.

3.3 Speech-language proficiency assessment

We utilized animal picture and letter boards for language training, featuring ten different animals: bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, and giraffe (as seen in Fig. 8). The flowchart for the MSAP's language proficiency assessment is illustrated in Fig. 7. Graduate students and undergraduate research participants from the laboratory served as test subjects. A total of 10 native MAN speakers (5 females and 5 males) were selected to undergo speech recording, language training, and language proficiency assessment. The feasibility of the VSLP developed in this study was tested and validated. The step-by-step speech-language proficiency assessment is shown below.

Step 1) Randomly select an animal picture and letter board (as seen in Table 1).
→ **Step 2)** The participant views the animal (object) and captures its image.
→ **Step 3)** Automatic OD and OC using "YOLOv9" on VSLP.
→ **Step 4)** On the basis of the recognized animal class, the VSLP generates a corresponding short phrase construction (as seen in Table 1).
→ **Step 5)** The generated phrase is displayed on the MR glasses.
→ **Step 6)** The participant reads the phrase aloud while the VSLP records the speech.
→ **Step 7)** Identify the participant's gender and nationality using MFCC + 1D-CNN.
→ **Step 8**) On the basis of the participant's gender and nationality, generate reference speech signals using GoogleTS + Python Playsound.
→ **Step 9)** Evaluate the participant's speech-language proficiency.
→ **Step 10)** The assessment score is displayed on the MR glasses.
→ **Step 11)** The VSLP displays the mispronounced or omitted words on the MR glasses.
→ **Step 12)** The reference speech signal is played three times, allowing the participant to repeat training through the MR interactive interface.
→ **Step 13)** Repeat until mastery is achieved: repeat Step 6 to Step 12, until the participant achieves a score of 80% or higher (as indicated in Table 9).
→ **Step 14)** A new random picture and letter board is generated. Repeat Step 2 to Step 13.

Table 7
Tenfold cross-validation results.

Fold	1	2	3	4	5	6	7	8	9	10
Loss	0.0181	0.0038	0.0142	0.0015	0.0059	0.0052	0.0177	0.0081	0.0154	0.0123
Accuracy (%)	98.76	99.75	99.50	99.75	99.75	99.25	99.75	99.25	99.48	99.25

Table 8
Tenfold cross-validation results for classifier performance evaluation.

Fold	*Precision* (%)	*Recall* (%)	*F*1 *score*	*Accuracy* (%)
1	99.75	99.75	0.9975	99.75
2	99.75	99.75	0.9975	99.75
3	99.75	99.75	0.9975	99.75
4	98.79	98.75	0.9875	98.75
5	100.00	100.00	1.0000	100.00
6	99.50	99.50	0.9949	99.50
7	99.75	99.75	0.9975	99.75
8	99.26	99.25	0.9925	99.25
9	99.51	99.50	0.9950	99.50
10	100.00	100.00	1.0000	100.00
Average	99.60	99.60	0.9959	99.60

Table 9
Experimental results of speech-language proficiency assessment for MSAPs.

Testing no.	Animal category	Reference short phrase sentence	Participant's spoken short phrase sentence	Rating	Score (%)
1	Cow	A cow is grazing.. (一隻牛正在吃草)	A … (一 …)	1	14.29
2			A … (一隻 …)	2	28.75
3			A Cow … (一隻牛 …)	3	42.86
4			Grazing A Cow… (吃草一隻牛)	2	28.75
1	Bird	This is a beautiful bird. (這是一隻美麗的鳥)	This … (這 …)	1	12.50
2			This Is… (這是 … 這 …)	2	25.00
3			… A Beautiful … Bird (… 一隻美麗 … 鳥)	4	75.00
4			This Is A Beautiful … (這是一隻美 …)	4	75.00
1	Elephant	This is a strong elephant. (這是一隻強壯的大象)	This Is A … (這是一 …)	2	33.33
2			This Is A Strong … (這是一隻強壯 …)	4	66.67
3			This … Strong … Elephant (這是…強壯…大象)	3	55.56
4			This is a strong elephant. (這是一隻強壯的大象)	5	100.00
1	Giraffe	A giraffe has a very long neck. (長頸鹿脖子很長)	…	1	0.00
2			… Giraffe… (長頸鹿 …)	2	20.00
3			… Giraffe…Long… (長頸 … 長)	2	20.00
4			…Giraffe…Neck... Very Long (長頸鹿脖子很長)	5	100.00

We successfully developed a rehabilitation assistive tool designed to provide personalized operations, voice prompts, and scenario-based speech-language training. This allows participants (MSAPs) to repeatedly practice daily under specific guidance at home. As shown in Table 9, the experimental results of the speech-language proficiency assessment indicate that the participant's scores improve over time with repeated practice. Additionally, through the repetition process, participants can identify areas where they struggle or have vocabulary omissions. By following the standardized training process (shown in Fig. 7), MSAPs can undergo intensive, personalized, and repetitive training without requiring the presence of an SLP. The proposed VSLP effectively enhances rehabilitation outcomes, including improvements in auditory comprehension, expressive abilities, and overall communication skills. In addition, for OD applications, YOLOv9 incorporates an enhanced network configuration designed to improve the detection *accuracy* for small and overlapping objects, which is particularly

beneficial for fine-grained recognition tasks, such as distinguishing various animal cards. Its model has a relatively low computational module, enabling deployment on devices with limited hardware resources, such as AR/MR glasses or mobile platforms.

4. Conclusions

Aphasia is a condition caused by brain damage and leads to impairments in language ability and comprehension. Aphasia patients may experience varying degrees of difficulty in auditory comprehension, verbal expression, reading comprehension, and writing. Clinically, the earlier aphasia is treated, the better the recovery effect. Patients also need daily practice, usually guided by a SLP, to achieve rehabilitation goals. Speech-language training treatment is currently the most direct clinical rehabilitation manner for aphasia patients. In addition to one-on-one communication interactions, SLPs assess the effectiveness of a patient's speech-language recovery on the basis of individual cases. However, this manner faces significant challenges, including limited manpower and time-consuming procedures. According to statistics from the Ministry of Health and Welfare, Taiwan, the first six months are considered to be the golden period for rehabilitation. Care services face a significant shortage and uneven distribution of SLPs, with only 4.2 available per 100 thousand people. These SLPs are responsible for treating children with speech-language delay, selective mutism, adults with speech-language impairment, language-based learning disability, and aphasia. In Taiwan, there are at least 65 thousand aphasia patients. Because of the limited availability of SLPs in care services, the demand for speech therapy far exceeds the supply. Even if each SLP can dedicate two days per week to speech-language rehabilitation programs, the demand remains unmet. We proposed the development of a rehabilitation assistive tool for MSAPs, integrating MR glasses, audiovisual processing units, and DL-based algorithms, the so-called VSLP. The VSLP has been designed to support aphasia rehabilitation, enabling patients to practice independently within their home environment and daily routines. By utilizing MR technology to create interactive scenarios, patients can actively engage in self-directed rehabilitation and effectively achieve their training goals. Additionally, the MR-based assistive tool enables the automatic assessment of a patient's language recovery progress throughout the rehabilitation process. The visuals and speech recognition have become key functions in an intelligent assistive system. In this study, OD, OC, and ASR technologies were combined to develop an assistive tool applicable to multilingual learning, language training, and language rehabilitation, and then integrated into an MR platform equipped with audio recording, speech feedback, and real-time visual feedback functionalities. During the language learning and rehabilitation, the proposed assistive tool can automatically recognize flashcards within the visual field, and then generate the corresponding learning sentences and standard pronunciations of the detected objects, and employ a word-by-word comparison mechanism to evaluate the *accuracy* of the user's pronunciation. Real-time inference is then performed to display the pronunciation similarity score on the visual feedback interface. Through repeated practice, this approach enhances the user's pronunciation *accuracy* and semantic comprehension, further promoting learning motivation. In addition, the assistive tool's

adaptability can be extended to various applications, including language disorder learning for young children, language instruction in special education, second language acquisition, and cognitive rehabilitation for older adults.

Acknowledgments

This work was supported by the National Science and Technology Council (NSTC), under contract number: NSTC 112-2635-E-167 -001, duration: August 1, 2023–October 31, 2024.

References

1 Anatomy & Physiology-Wernicke area and Broca area, Definition, Location, Function, & Facts (2025). https://www.britannica.com/science/Wernicke-area
2 Language in the Brain – Psychology of Language (2021). https://opentextbc.ca/psyclanguage/chapter/language-in-the-brain/
3 M. Devi Siva Rama Saran, B. K. Akshayaa, R. Sai Raghavendra, N. Poornima, and G. Jyothish Lal: 2024 15th Int. Conf. Computing Communication and Networking Technologies (ICCCNT) (Kamand, India, 2024) 1. https://doi.org/10.1109/ICCCNT61001.2024.10724563
4 M. F. Bonner, S. Ash, and M. Grossman: Curr. Neurol. Neuroscience Rep. **10** (2010) 484. https://doi.org/10.1007/s11910-010-0140-4
5 M. L. Gorno-Tempini, A. E. Hillis, S. Weintraub A. Kertesz, M. Mendez, S. F. Cappa, J. M. Ogar, J. D. Rohrer, S. Black, B. F. Boeve, F. Manes, N. F. Dronkers, R. Vandenberghe, K. Rascovsky, K. Patterson, B. L. Miller, D. S. Knopman, J. R. Hodges, M. M. Mesulam, and M. Grossman: Neurology **76** (2011) 1006. https://doi.org/10.1212/WNL.0b013e31821103e6
6 K. Hilari and S. Northcott: Aphasiology **20** (2006) 17. https://doi.org/10.1080/02687030500279982
7 S. Engelter, M. Gostynski, S. Papa, M. Frei, and C. Born: Stroke **37** (2006) 1379. https://doi.org/10.1161/01.STR.0000221815.64093.8c
8 L. R. Cherney, J. P. Patterson, A. M. Raymer, T. Frymark, and T. Schooling: J. Speech Lang. Hear. Res. **5** (2008) 1282. https://doi.org/10.1044/1092-4388(2008/07-0206)
9 R. Patel, D. Chandalia, A. Nayak, A. Jeyabose, and D. Jijo: IEEE Access **13** (2025) 61192. https://doi.org/10.1109/ACCESS.2025.3553530
10 M. W. Khan, M. S. Obaidat, K. Mahmood, B. Sadoun, H. Muhammad, S. Badar, and W. Gao: IEEE Internet Things J. **12** (2025)17649. https://doi.org/10.1109/JIOT.2025.3537640
11 M. A. Uddin, R. K. Pathan, M. S. Hossain, and M. Biswas: J. Inf. Telecommun. **6** (2022) 27. https://doi.org/10.1080/24751839.2021.1983318
12 G. Li, Y. Liu, and X. Wang: 2023 IEEE 5th Int. Conf. Civil Aviation Safety and Information Technology (ICCASIT) (Dali, China, 2023) 956. https://doi.org/10.1109/ICCASIT58768.2023.10351697
13 H.-Y. Lai, C.-C. Hu, C.-H. Wen, J.-X. Wu, N.-S. Pai, C.-Y. Yeh, and C.-H. Lin: IEEE Access **12** (2024) 102962. https://doi.org/10.1109/ACCESS.2024.3430296
14 F. Li, M. Liu, Y. Zhao, L. Kong, L. Dong, X. Liu, and M. Hu: EURASIP J. Adv. Signal Process. **59** (2019) 1. https://doi.org/10.1186/s13634-019-0651-3
15 Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang, Z. Zhou, B. Li, M. Ma, W. Chan, J. Yu, Y. Wang, L. Cao, K. C. Sim, B. Ramabhadran, T. N. Sainath, F. Beaufays, Z. Chen, Q. V. Le, C.-C. Chiu, R. Pang, and Y. Wuet: IEEE J. Sel. Top. Signal Process. **16** (2022) 1519. https://doi.org/10.48550/arXiv.2109.13226
16 S.-A. Li, Y.-Yi. Liu, Y.-C. Chen, H.-M. Feng, P.-K. Shen, and Y.-C. Wu: Appl. Sci. **13** (2023) 1. https://doi.org/10.3390/app13053359
17 Common Objects in Context (2025). https://cocodataset.org/#home
18 Open Speech and Language Resources, MAGICDATA Mandarin Chinese Read Speech Corpus (2025). https://www.openslr.org/68
19 The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) Training and Test Data (2023). https://goo.gl/l0sPwz
20 CLDPS (Corpora e Lessici dell' Italiano Parlato e Scritto (2024). http://www.clips.unina.it/en/index.js

21 ArabiCorpus: The Quranic Arabic Corpus - Word by Word Grammar, Syntax and Morphology of the Holy Quran (2017). https://corpus.quran.com
22 C.-H. Lin, H-Y. Lai, P.-T. Huang, P.-Y. Chen, and C.-M. Li: IET Signal Process. **17** (2023) 1. https://doi.org/10.1049/sil2.12216
23 C. Arhiliuc, J. Mitrovic, and M. Granitzer: Proc. 12th Conf. Language Resources and Evaluation (LREC 2020) (Marseille, France, 2020) 5624–5630. https://aclanthology.org/2020.lrec-1.690/
24 S. C. Purdy, I. Wanigasekara, O. M. Cañete, C. Moore, and C. M. McCann: Semin. Hear. **37** (2016) 233. https://doi.org/10.1055/s-0036-1584408
25 D. Kadojic, B. R. Bijelic, R. Radanovic, M. Porobic, J. Rimac, and M. Dikanovi: Acta Clin Croat. **51** (2012) 221. PMID: 23115946
26 American Speech-Language-Hearing Association (ASLHA), Augmentative and Alternative Communication (AAC). https://www.asha.org/practice-portal/professional-issues/augmentative-and-alternativecommunication/?srsltid=AfmBOorxWootHKgRkBvRM0RogojfuYkn_m4ENVZjzztiCxf_1QWDQWXW#collapse_1
27 S. S. Mahmoud, A. Kumar, Y. Tang, Y. Li, X. Gu, J. Fu, and Q. Fang: IEEE J. Biomed. Health Info. **24** (2020) 3191. https://doi.org/10.1109/JBHI.2020.3011104
28 Y. Elsahar, S. Hu, K. Bouazza-Marouf, D. Kerr, and A. Mansor: Sensors **19** (2019) 1. https://doi.org/10.3390/s19081911
29 N. Iftikhar, A. M. Naz, H. M. T. Zia, M. B. Bhatti, and Asia: J. Popul. Ther. Clin. Pharlmacol. **30** (2023) 2254. https://doi.org/10.53555/jptcp.v30i18.3432
30 J. Wanga, N. Wagleya, M. L. Riceb, and J. R. Bootha: Cortex **6** (2021) 169. https://doi.org/10.1016/j.cortex.2021.09.006
31 J. Zou and H. Wang: IEEE Access **12** (2024) 124160. https://doi.org/10.1109/ACCESS.2024.3453931
32 J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González: Mach. Learn. Knowl. Extr. **5** (2023) 1680. https://doi.org/10.3390/make5040083
33 F. Li, M. Liu, Y. Zhao, L. Kong, L. Dong, X. Liu, and M. Hu: EURASIP J. Adv. Signal Process. **59** (2019) 1. https://doi.org/10.1186/s13634-019-0651-3
34 K. Kondo: Signals and Communication Technology (Springer, Japan, 2012). https://www.springer.com/series/4748
35 W. Lin, D. Tao, J. Kacprzyk, Z. Li, E. Izquierdo, and H. Wang: Studies in Computational Intelligence - Speech Quality Assessment 346 (Springer, Verlag Berlin, 2011) pp. 623–654. https://doi.org/0.1007/978-3-642-19551-8
36 S. H. K. Chen, M. R. McNeil, and S. R. Pratt: Speech Lang. Hear. **16** (2013) 37. https://doi.org/10.1179/2050571X12Z.0000000006.
37 J. Allen, H. Liu, S. Iqbal, D. Zheng, and G. Stansby: Physiol. Meas. **42** (2021) 1. https://doi.org/10.1088/1361-6579/abf9f3
38 J. J. Lee, J. H. Heo, J. H. Han, B. R. Kim, H. Y. Gwon, and Y. R. Yoon: J. Med. Biol. Eng. **40** (2020) 282. https://doi.org/10.1007/s40846-020-00507-w
39 M. Panwar, A. Gautam, R. Dutt, and A. Acharyya: 2020 IEEE Int. Symp. Circuits and Systems (Seville, Spain, 2020) 1. https://doi.org/10.1109/ISCAS45731.2020.9180636
40 D. P. Kingma and J. L. Ba: 3rd Int. Conf. Learning Representations (San Diego, 2015). https://hdl.handle.net/11245/1.505367
41 National Academy for Educational Research, Mandarin Chinese Corpus (2024). https://coct.naer.edu.tw/?utm_source=chatgpt.com
42 C.-C. Hu: Thesis for Degree of Master, Department of Electrical Engineering, National Chin-Yi University of Technology (July, 2025).