2024 33rd IEEE International Conference on
Robot and Human Interactive Communication (RO-MAN)
August 26-30, 2024. Pasadena, California, USA.

Voice-Controlled Human-Machine Interface for an Assistive Exoskeleton Glove Aiding Patients with Brachial Plexus Injuries

Yunfei Guo[1], Wenda Xu[2], Cesar Bravo[3] and Pinhas Ben-Tzvi[4]

***Abstract*— This paper introduces a voice-controlled Human Machine Interface (HMI) tailored for an assistive robotic exoskeleton glove, aimed at assisting patients coping with Brachial Plexus Injuries (BPI) in regaining their lost grasping functionality. The development of this HMI draws upon clinical experimentation results, forming a foundation for its design. The paper delves into the challenges encountered while employing a prior voice-based HMI, which necessitated an internet connection for complex computations and exhibited limitations in effectively processing concise commands. To address these issues, an innovative voice-controlled HMI system is proposed, featuring fixed-word detection to replace the speech-to-text (STT) converter and the Neutral Language Processor (NLP) to reduce computational overhead. Furthermore, the new HMI replaces the previous text-independent speaker verification with a text-dependent, one-shot learning approach. This enhancement streamlines custom retraining, significantly improving speaker verification accuracy for concise commands. Experimental results substantiate the applicability of the proposed voice-controlled HMI for assisting individuals with BPI through specialized exoskeleton gloves.**

***Index Terms*— Assistive Robotics, Exoskeleton Glove, Voice-controlled HMI, Wearable Robotics, One-shot Learning**

I. INTRODUCTION

A. Assistive Robotic Exoskeleton Gloves

Assistive robotic exoskeleton gloves have gained widespread use in postsurgical physical therapy for Brachial Plexus Injuries (BPI) [1]–[5]. BPI, which is typically the result of motorcycle or snowmobile accidents, inflicts damage to the neural system of the hand, arm, and shoulder, leading to compromised mobility and sensation. Although surgical interventions can partially restore arm and shoulder function, they often do not address hand-related problems, making exoskeleton gloves a promising method to prevent hand muscle atrophy during therapy [6].

Beyond physical therapy, wearable robotic exoskeleton gloves extend their utility to assist patients with BPI in everyday activities [7], [8].

B. Exoskeleton Glove Human Machine Interface for Patients with BPI

The Human-Machine Interface (HMI) plays a pivotal role in enabling users to control exoskeleton gloves with minimal effort. Unlike patients recovering from post-stroke symptoms, those with BPI often lack control over their muscles on the paralyzed hand and arm. During physical therapy of the hand muscle, patients use the healthy arm to assist the paralyzed hand and arm. The physical therapy procedure makes it difficult to find a location to place electromyography (EMG) sensors, making EMG-based HMIs unsuitable for their needs [9]–[12].

Although noninvasive electroencephalogram (EEG)-based HMIs have been explored, they require the use of EEG probes or headsets, which are less cost-effective and portable compared to voice-based alternatives [13]–[15].

In contrast, voice-controlled HMIs offer exceptional wearability and robust intention detection, making them a preferred choice for assistive exoskeleton gloves [16]–[18]. Furthermore, voice-based HMIs equipped with speaker verification ensure operational safety for assistive exoskeletons, making them the focus of this research.

This paper introduces a voice-based HMI design for an assistive robotic exoskeleton glove, addressing speaker verification challenges. This work is part of a broader effort to develop a state-of-the-art exoskeleton glove system. Initially, we designed an assistive exoskeleton glove paired with a voice-controlled HMI and conducted clinical experiments. Section II-A describes the exoskeleton glove used in these experiments. Section II-B discusses the initial voice-controlled HMI, while Section III presents the clinical results. Based on these results, the exoskeleton glove and HMI were modified and improved [19]. Sections IV and V detail the revised HMI, and Section VI covers the updated system and experiments.

II. RELATED WORK

A. Assistive Exoskeleton Glove

This study used a linkage-driven exoskeleton glove with 7 degrees of freedom (DOF) in the clinical experiment [2]. Its design incorporates Series Elastic Actuators (SEAs) to empower the movement of each finger, wrist, and thumb thenar, while simultaneously offering force sensing capabilities at each linkage endpoint. This exoskeleton facilitates five

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R21HD095027. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

This material is partly based upon work supported by (while serving at) the National Science Foundation. *(Corresponding author: Pinhas Ben-Tzvi)*

[1] Yunfei Guo is with the Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: yunfei96@vt.edu).

[2] Wenda Xu is with the Department of Mechanical Engineering, Virginia Tech, Blacksburg, VA, 24061 USA (e-mail: wenda@vt.edu).

[3] Cesar Bravo is with the Carilion Clinic Institute of Orthopaedics and Neurosciences, Virginia Tech Carilion School of Medicine, Roanoke, VA 24016 USA (e-mail: cjbravo@carilionclinic.org).

[4] Pinhas Ben-Tzvi is with the Departments of Mechanical Engineering and Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, 24061 USA (e-mail: bentzvi@vt.edu).

979-8-3503-7502-2/24/$31.00 ©2024 IEEE

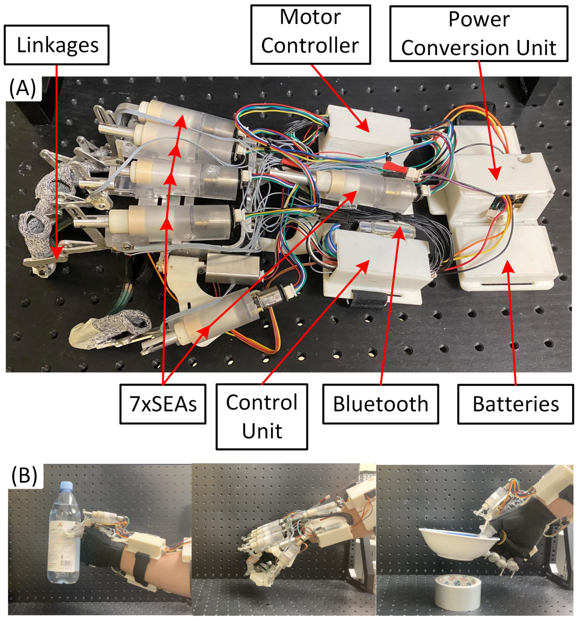

Fig. 1. The assistive exoskeleton glove used in the clinical experiment. (A) An integrated prototype of the glove. (B) Examples of grasping experiments with the glove.

fundamental grasping types, including cylinder grasp, sphere grasp, tripod grasp, tip grasp, and lateral grasp.

The exoskeleton glove contains an on-board microcontroller, batteries, and Bluetooth connectivity, enabling wireless operation for approximately 2.5 hours. The entire glove weighs 759 grams, including batteries and control units. A visual representation of the integrated exoskeleton glove prototype is shown in Fig. 1.

B. Voice-controlled HMI with Text-independent Speaker Verification

The exoskeleton glove is equipped with a voice-controlled Human-Machine Interface (HMI) featuring an embedded text-independent speaker verification component [18], [20]. Leveraging a Bluetooth Earpod as the voice input device, this HMI responds to personalized voice commands, rendering interaction effortless.

Upon activation, users communicate with the HMI via voice, which is subsequently transformed into text through Google's online speech-to-text (STT) API. Keyword analysis is performed to discern the intended grasp type. If the voice command proves to be valid, we employ a text-independent deep learning-based speaker verification process. This technique used the VoxCeleb dataset for training a speaker's utterance extractor, gauging the cosine distance between the incoming speaker's utterance and the enrolled user. Any similarity score that falls below the predefined threshold results in rejection, ensuring exclusive control of the exoskeleton glove by the enrolled user. The architectural layout of this voice-based HMI is outlined in Fig. 2.

This method has achieved an Equal Error Rate (EER) of only 12.4% in the VoxCeleb1 validation dataset. The HMI system operates with minimal latency, executing efficiently on a single-thread Intel i7-8750H processor with 2GB of RAM. Its performance is underlined by an average accuracy rate of 91.4% in correctly classifying and verifying voice commands [18], [20].

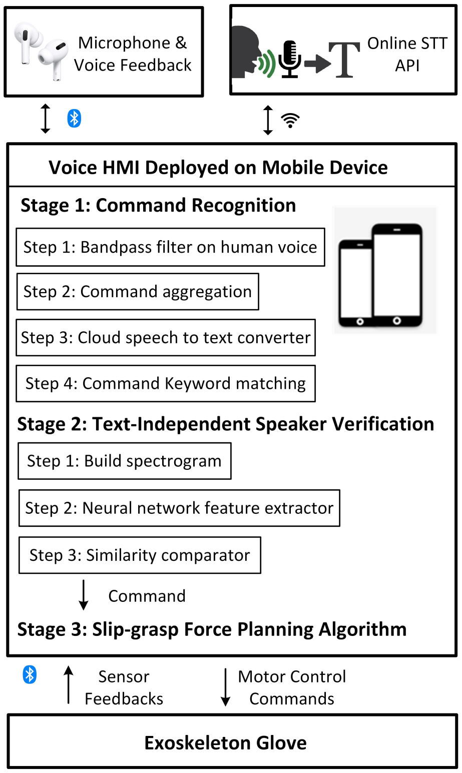
Microphone & Voice Feedback
Online STT API
Voice HMI Deployed on Mobile Device
Stage 1: Command Recognition
Step 1: Bandpass filter on human voice
Step 2: Command aggregation
Step 3: Cloud speech to text converter
Step 4: Command Keyword matching
Stage 2: Text-Independent Speaker Verification
Step 1: Build spectrogram
Step 2: Neural network feature extractor
Step 3: Similarity comparator
Command
Stage 3: Slip-grasp Force Planning Algorithm
Sensor Feedbacks
Motor Control Commands
Exoskeleton Glove

Fig. 2. The structure of voice-based HMI used in clinical experiments

III. Clinical Experiment Challenges

In close collaboration with the Carilion Clinic, a series of clinical experiments were carried out under the IRB-19-330 protocol, involving patients affected by Brachial Plexus Injuries (BPI). In four clinical trials involving three participants, considerable success was achieved in restoring grasping ability in three of the trials.

Fig. 3(A) portrays a BPI-afflicted subject with muscle atrophy in her right hand, despite undergoing surgery for nerve system reconstruction in her shoulder and forearm. Notably, her shoulder and forearm muscles remained weak. As evidenced in Fig. 3(B)-(E), she had to rely on her unaffected hand to support her paralyzed arm and hand when attempting to grasp objects. The application of the voice-controlled assistive exoskeleton glove, featuring automatic force planning, proved instrumental in partially reinstating her hand's grasping capabilities, as demonstrated in Fig. 3(C) and (E).

All subjects effectively operated the voice-controlled HMI during clinical experiments. However, this HMI revealed two practical challenges. First, it relied on an internet connection for speech-to-text (STT) conversion and Natural Language Processing (NLP) for voice command recognition. Consequently, when operating in areas with limited Internet access, the HMI experienced significant latency, reaching up to

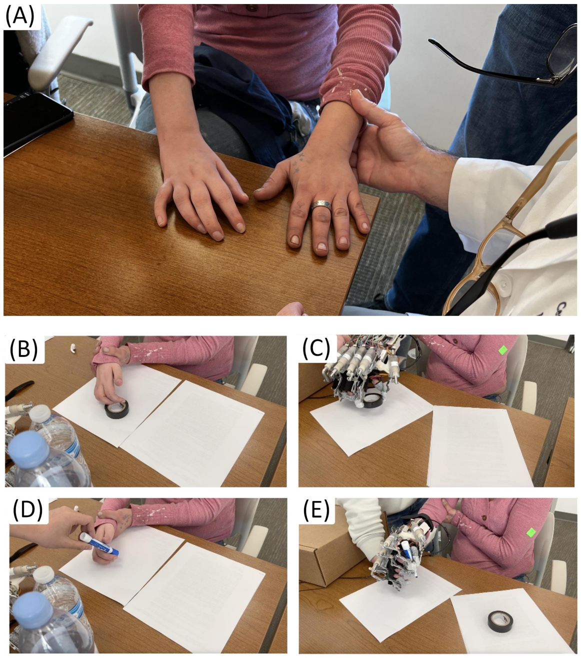

Fig. 3. Clinical experiment performed using the assistive exoskeleton glove. (A) Patient with BPI on her right hand. (B) The patient failed to grasp a duct tape. (C) The patient successfully grasped the duct tape with the help of the exoskeleton glove. (D) The patient failed to grasp a marker.(E) The patient successfully grasped the marker with the help of the exoskeleton glove.

700ms. In particular, two out of three subjects deemed this latency higher than desirable, suggesting that the processing time should be reduced by half.

Second, the voice commands typically used were concise, leading to suboptimal performance in text-independent speaker verification, yielding a 22% Equal Error Rate (EER) compared to the more favorable 12.4% EER achieved in the VoxCeleb1 validation dataset. The core algorithm, reliant on deep learning for generalized feature extraction, presented challenges when dealing with edge cases. For instance, one subject's mother could effortlessly activate the exoskeleton glove, which was initially enrolled using the subject's voice. Retraining the model for each subject using the VoxCeleb1 data set, which required more than 30 hours with a GTX Titan XP GPU, proved impractical.

In response to these challenges, vital modifications were made to the previous voice-controlled HMI. First, STT and NLP were replaced with a fixed command keyword detector, eliminating the need for an Internet connection. Second, the transition to using fixed commands to operate the exoskeleton glove enabled the implementation of a one-shot learning-based text-dependent speaker verification method. Subsequent sections offer detailed technical insights into these enhancements.

IV. Command Detection

In the previous voice HMI, STT and NLP methods were used to extract voice commands due to their robustness. For instance, phrases like "grasp a cup" or "grasp a water bottle" effectively triggered a cylinder grasp by the exoskeleton glove. However, based on valuable patient feedback, a more user-friendly approach was favored, allowing patients to directly choose the grasp type, typically involving shorter commands. Given that the exoskeleton used in this research accommodates five fundamental grasp types, users found it convenient to memorize these commands. To enhance the accuracy of fixed command detection, fixed commands were incorporated into the grammar of the STT and NLP models, effectively reducing the Word-Error-Rate (WER). This new approach was introduced midway through the clinical experiments and garnered positive feedback from patients.

Following the conclusion of the experiments, the need for employing the STT and NLP methods as a command detector became redundant due to the use of fixed commands. Researchers have explored various efficient approaches for fixed command detection, with two common methodologies prevailing. Some have advocated for a one-shot learning method, comparing the input voice command with enrolled commands [21]. Conversely, others have employed neural network methods as feature extractors, coupled with the Hidden Markov Model (HMM) approach to decipher the letter sequence of the input command [22], [23]. The field has witnessed extensive research efforts, resulting in the availability of multiple APIs. Notably, the Picovoice Porcupine wake word detection API has been favored over other commonly used APIs such as Pocketsphinx [24] or Snowboy [21], primarily due to its superior performance. This selection aligns with the API's minimal computational demands and its capacity to function without Internet connectivity.

V. Speaker Verification

Two prevalent text-dependent speaker verification approaches have been widely explored. The first approach embraces a one-shot learning methodology for text-dependent speaker verification [25], [26]. This approach offers the advantage of requiring significantly less data and training time, focusing solely on the development of a comparison network [27]. Nevertheless, it necessitates specific training data aligned with the verification command, which can be labor intensive to collect for each unique command.

In the second approach, some researchers have opted to split the text-dependent speaker verification task into two distinct tasks: Speech-to-Text (STT) and text-independent speaker verification. The text-independent speaker verification method leverages a Convolution Neural Network (CNN) feature extractor and a cosine distance comparator to differentiate between different speakers [28], [29]. This approach avoids the data availability issue by drawing from a text-independent speaker verification dataset like VoxCeleb. While theoretically more robust, practical performance is influenced by factors such as the length and type of voice commands, as well as the speaker's characteristics. This CNN-cosine comparator method was employed in the previous HMI.

To select the most suitable solution for this application, both methods are implemented and compared.

A. The Collected Speaker Verification Dataset

To evaluate the performance of both speaker verification methods, a dataset encompassing approximately 2000 voice commands from seven distinct subjects (referred to as Speaker A to G) was meticulously curated. It is noteworthy that all seven subjects share several common attributes: they are all males aged between 24 to 30, and they possess a Mandarin language background. Within this dataset, each subject contributed a set of five standardized commands, which include: “hey glove,” “cylinder grasp,” “tripod grasp,” “lateral grasp,” and “release object.”

In the case of the one-shot learning method, the one-shot comparison network was trained using data from five specific speakers (Speakers A to E), while the remaining two speakers (F and G) were designated for validation purposes. The validation process involved executing speaker verification for each command uttered by Speakers F and G against the corresponding command articulated by Speakers A to E.

B. One-shot Learning Based Text-Dependent Speaker Verification

The essence of one-shot learning lies in the creation of a comparison neural network responsible for binary classification based on the similarity between two inputs. The structural components of the one-shot learning method are illustrated in Fig. 4. At runtime, the first spectrogram represents the user input. The second spectrogram is chosen from a dictionary of spectrograms corresponding to each command and each user. The algorithm can be dissected into two fundamental segments: preprocessing and the comparison neural network. Below, a closer examination is conducted to elucidate the design choices underpinning this framework.

First, voice commands are recorded using a single channel microphone at a sampling frequency of 16,000 Hz. To standardize input data, a fixed length L is established for each valid voice command, depending on the length of the command itself. If the input voice command exceeds this predetermined length, it is trimmed at both ends. Conversely, if the input voice command falls short of L, zero-padding is employed to fill the void. To enable input compatibility with convolutional neural networks, the data undergoes a conversion process into a spectrum. Two widely adopted audio preprocessing conversion methods were evaluated: Mel Frequency Cepstral Coefficients (MFCC) and Mel-spectrum (Mel). Both methods were rigorously assessed on speaker verification validation tasks, with performance gauged using the Equal Error Rate (EER). After meticulous tuning of the number of coefficients and Mel bands, the optimal results are presented in Tab. I. It was discerned that the Mel-spectrum method outperformed the MFCC method.

Second, the performance of several commonly used comparator networks was analyzed, including VGG-16, MobileNet V2 (MBN V2), and Resnet-50 (Res-50). The results are summarized in Tab. I. It was evident that the VGG-16 network achieved the lowest EER, establishing its superiority in this context.

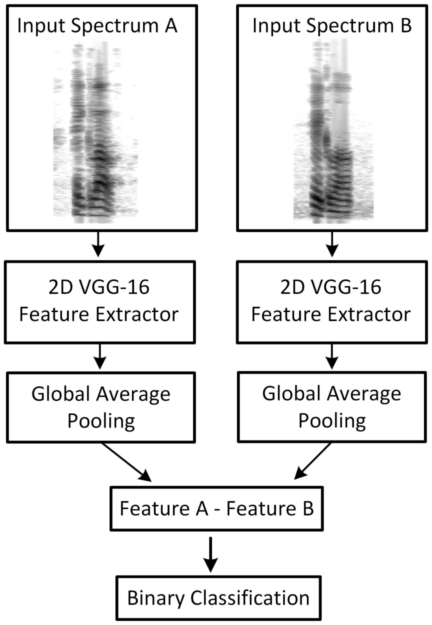

Fig. 4. The structure of the one-shot learning based text-dependent speaker verification method. Spectrum A: the spectrum of the user’s input during runtime for the voice command ”release object.” Spectrum B: the stored spectrum of the same user’s ”release object” voice command.

TABLE I

COMPARISON OF DIFFERENT ONE-SHOT LEARNING SPEAKER VERIFICATION DESIGN CHOICES

Preprocess	Comparator Networks	EER
MFCC	VGG-16	34%
Mel	VGG-16	23%
Mel	MBN V2	28%
Mel	Res-50	23%

C. Speaker Verification Comparison

This subsection evaluated the effectiveness of text-dependent and text-independent speaker verification by measuring their performance using the Equal Error Rate (EER). A lower EER indicates a more accurate speaker verification. Fig. 5 offers a comprehensive view of the speaker verification EER for speaker F and G across two distinct commands. These specific commands have been chosen as illustrative examples to underscore the impact of command type and speaker variability on EER. Notably, when evaluating all commands for speaker F and G against speaker A to E, both CNN-Cosine and One-shot methods exhibit a similar average EER, as indicated in Tab. II. It is difficult to come to a definite conclusion regarding which method is more effective due to the restricted amount of data available.

However, it is worth highlighting that the one-shot method emerges as the more suitable choice for this application because of the following three reasons. First, it achieved comparable performance to the CNN-cosine comparator method, while demanding significantly fewer data and shorter training

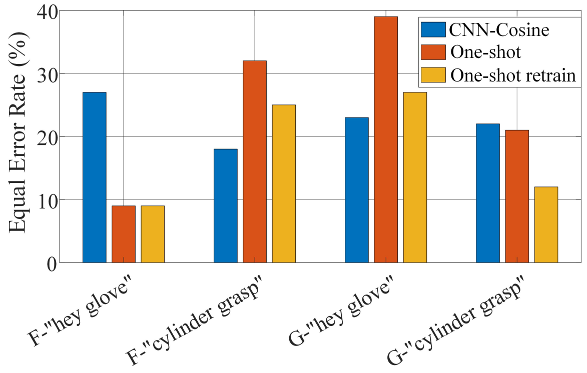

Fig. 5. Comprehensive view of the speaker verification Equal Error Rate for speaker F and G across two distinct commands. F-"hey glove": Speaker verification was conducted on the "hey glove" command uttered by speaker F compared to the same command spoken by the remaining speakers.

TABLE II
PERFORMANCE COMPARISON BETWEEN CNN-COSINE AND ONE-SHOT SPEAKER VERIFICATION METHODS

Method	Command	Task: speaker x vs. x	EER
CNN-Cosine	All	F, G vs. A-E	22%
One-shot	All	F, G vs. A-E	23%
One-shot retrain	All	F, G vs. A-E	**16%**

times, as evidenced in Tab. III. Additionally, inference times were measured by running both methods on an Intel E5-1260 CPU, while training times were determined by executing both methods on an NVIDIA GTX Titan XP GPU. Furthermore, exploiting the rapid retraining capability of the one-shot method, which allows the inclusion of one of the held-out speakers for testing against other speakers, can enhance speaker verification EER on the collected dataset. As shown in Tab. II, this enhancement leads to an improvement from 23% to 16% in the EER.

TABLE III
SPEAKER VERIFICATION COMPARISON

Method	CNN-Cosine	One-shot
Inference Speed	∼94ms	∼190ms
Training Time	30+hr	∼20min
Training Data Size	300,000+ utterances	∼2000 utterances

VI. IMPROVED HUMAN MACHINE INTERFACE

This section discusses the structure of the improved HMI. The voice-controlled HMI was modified based on the previously mentioned discoveries. The design of the new voice HMI is shown in Fig. 6. It functions by capturing input from the microphone, recognizing the command through the Picovoice Porcupine wake-word API, and subsequently transmitting the recognized command to a one-shot learning speaker verification model.

A comparison of the processing speed between the previous and new HMIs is detailed in Tab. IV. The inference time was evaluated by executing both methods on an Intel E5-1260 CPU. It is noteworthy that the one-shot method was marginally slower by approximately 100 ms compared to the previous method, given the assumption of an ideal Internet connection. This discrepancy can be attributed to the fact that all computations were executed locally.

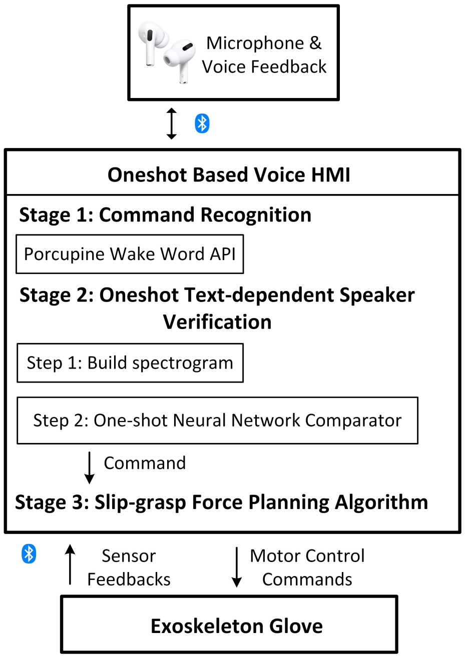
Microphone &
Voice Feedback
Oneshot Based Voice HMI
Stage 1: Command Recognition
Porcupine Wake Word API
Stage 2: Oneshot Text-dependent Speaker Verification
Step 1: Build spectrogram
Step 2: One-shot Neural Network Comparator
Command
Stage 3: Slip-grasp Force Planning Algorithm
Sensor Feedbacks
Motor Control Commands
Exoskeleton Glove

Fig. 6. The structure of proposed voice-controlled HMI using the one-shot learning method.

TABLE IV
TIME COST COMPARISON TO PROCESS ONE COMMAND

	Google API+CNN-Cosine	Porcupine + One-shot
CR	58+ ms	∼110ms
SV	∼94ms	∼190ms
Total	152+ms	∼300ms

CR: Command Recognition; SV: Speaker Verification

In addition, an experimental assessment was conducted on the complete HMI system, involving 200 voice commands sourced from two human subjects. This data set encompassed five types of command, each type comprising 20 trials per subject. Approximately 50% of the data was allocated to retraining the model. The resulting performance was tested against the previous HMI and other state-of-the-art HMIs, as illustrated in Tab. V. The proposed HMI was tested with 200 voice commands at a binary output threshold of 0.5. It achieved a classification accuracy of 98% and a verification true acceptance rate of 96.5%. The overall verification success rate was 94.5%.

VII. CONCLUSION

This paper proposed a voice-controlled human-machine interface (HMI) with speaker verification, specifically designed to assist patients with Brachial Plexus Injuries. The proposed HMI used a one-shot learning method to perform text-dependent speaker verification without the need for

TABLE V

COMPARISON BETWEEN THE PROPOSED HMI SYSTEM AND OTHER STATE-OF-THE-ART VOICE HMI SYSTEMS

Author	Method	Acc*	SV
Proposed	Porcupine + one-shot	94.5%	Yes
Yunfei, et al. [18] (old HMI)	GoogleAPI+CNN	91.4%	Yes
He, et al. [30]	GoogleAPI	92%	No
El-emary, et al. [31]	GMM	<85%	No
Gomez, et al. [32]	MG GMM+SM	88%	No
Gomez, et al. [32]	HMM	100%	No
Megalingam, et al. [33]	PocketSphinx:HMM	90%	No
Pleva, et al. [34]	Julius: HMM	91%	No
Guo, et al. [35]	LD3320 speech chip	94%	No

SV: speaker verification

Acc*: for HMIs without speaker verification, Acc is the command classification accuracy. For this paper, Acc stands for successful rate, which is the command classification accuracy times the verification true acceptance rate.

GMM: Gaussian Mixture Model

MG GMM+SM: Mouth gesture based detection using GMM and state machine

HMM: Hidden Markov Model

an Internet connection. Compared to the previous voice-controlled HMI, this has also drastically reduced training time and requires significantly less data during training.

The proposed Human-Machine Interface (HMI) achieved a 23% Equal Error Rate (EER) on speaker verification, which is similar to the performance of the prior HMI that employed a CNN-cosine comparator technique. In particular, the one-shot learning approach can be readily applied in real-world scenarios to retrain the neural network with user-specific voice commands. This technique yielded an reduction in EER from 23% to 16%, surpassing the CNN-cosine method by 6%.

In contrast to other state-of-the-art voice HMIs, the proposed interface has a 94.5% success rate in recognizing and verifying input commands. The HMI's response time delay on an E5-1260 CPU, approximately 300ms, was deemed acceptable by human subjects.

The comparative results underscore the advantages of the proposed method over other state-of-the-art voice HMIs, particularly in applications that require customized and concise commands. It is worth noting that this paper introduces preliminary findings on a voice control method utilizing one-shot learning. Further validation of its performance will require a larger training and validation dataset in the future. Furthermore, conducting more clinical experiments to gather user feedback will be essential for a comprehensive evaluation.

REFERENCES

[1] Q. A. Boser, M. R. Dawson, J. S. Schofield, G. Y. Dziwenko, and J. S. Hebert, "Defining the design requirements for an assistive powered hand exoskeleton: A pilot explorative interview study and case series," *Prosthetics and Orthotics International*, p. 0309364620963943, 2020.

[2] W. Xu, Y. Guo, C. Bravo, and P. Ben-Tzvi, "Design, control, and experimental evaluation of a novel robotic glove system for patients with brachial plexus injuries," *IEEE Transactions on Robotics*, vol. 39, no. 2, pp. 1637–1652, 2023.

[3] W. Xu, Y. Liu, and P. Ben-Tzvi, "Development of a novel low-profile robotic exoskeleton glove for patients with brachial plexus injuries," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 11 121–11 126.

[4] E. K. Jian, D. Gouwanda, T. K. Kheng *et al.*, "Wearable hand exoskeleton for activities of daily living," in *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE, 2018, pp. 221–225.

[5] P. Tran, D. Elliott, K. Herrin, and J. P. Desai, "Towards comprehensive evaluation of the flexotendon glove-iii: a case series evaluation in pediatric clinical cases and able-bodied adults," *Biomedical Engineering Letters*, pp. 1–10, 2023.

[6] N. Smania, G. Berto, E. La Marchina, C. Melotti, A. Midiri, L. Roncari, A. Zenorini, P. Ianes, A. Picelli, A. Waldner *et al.*, "Rehabilitation of brachial plexus injuries in adults and children," *Eur J Phys Rehabil Med*, vol. 48, no. 3, pp. 483–506, 2012.

[7] T. Bützer, O. Lambercy, J. Arata, and R. Gassert, "Fully wearable actuated soft exoskeleton for grasping assistance in everyday activities," *Soft robotics*, vol. 8, no. 2, pp. 128–143, 2021.

[8] Y. Chen, Z. Yang, and Y. Wen, "A soft exoskeleton glove for hand bilateral training via surface emg," *Sensors*, vol. 21, no. 2, p. 578, 2021.

[9] K. O. Thielbar, K. M. Triandafilou, H. C. Fischer, J. M. O'Toole, M. L. Corrigan, J. M. Ochoa, M. E. Stoykov, and D. G. Kamper, "Benefits of using a voice and emg-driven actuated glove to support occupational therapy for stroke survivors," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 3, pp. 297–305, 2017.

[10] Y. Chen, Z. Yang, and Y. Wen, "A soft exoskeleton glove for hand bilateral training via surface emg," *Sensors*, vol. 21, no. 2, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/2/578

[11] S. Cheon, D. Kim, S. Kim, B. Kang, J. Lee, H. Gong, S. Jo, K.-J. Cho, and J. Ahn, "Single emg sensor-driven robotic glove control for reliable augmentation of power grasping," *IEEE Transactions on Medical Robotics and Bionics*, vol. PP, pp. 1–1, 12 2020.

[12] K. Li, Z. Li, H. Zeng, and N. Wei, "Control of newly-designed wearable robotic hand exoskeleton based on surface electromyographic signals," *Frontiers in Neurorobotics*, vol. 15, p. 121, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbot.2021.711047

[13] L. Randazzo, I. Iturrate, S. Perdikis, and J. d. R. Millán, "mano: A wearable hand exoskeleton for activities of daily living and neurorehabilitation," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 500–507, 2017.

[14] C. E. Bouton, A. Shaikhouni, N. V. Annetta, M. A. Bockbrader, D. A. Friedenberg, D. M. Nielson, G. Sharma, P. B. Sederberg, B. C. Glenn, W. J. Mysiw *et al.*, "Restoring cortical control of functional movement in a human with quadriplegia," *Nature*, vol. 533, no. 7602, pp. 247–250, 2016.

[15] S. R. Soekadar, M. Witkowski, N. Vitiello, and N. Birbaumer, "An eeg/eog-based hybrid brain-neural computer interaction (bnci) system to control an exoskeleton for the paralyzed hand," *Biomedical Engineering/Biomedizinische Technik*, vol. 60, no. 3, pp. 199–205, 2015.

[16] P. Tran, S. Jeong, and J. P. Desai, "Voice-controlled flexible exotendon (flexotendon) glove for hand rehabilitation," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4834–4839.

[17] X. Wang, P. Tran, S. M. Callahan, S. L. Wolf, and J. P. Desai, "Towards the development of a voice-controlled exoskeleton system for restoring hand function," in *2019 International Symposium on Medical Robotics (ISMR)*, 2019, pp. 1–7.

[18] Y. Guo, W. Xu, S. Pradhan, C. Bravo, and P. Ben-Tzvi, "Personalized voice activated grasping system for a robotic exoskeleton glove," *Mechatronics*, vol. 83, p. 102745, 2022.

[19] W. Xu, Y. Guo, and P. Ben-Tzvi, "Robotic Exoskeleton Glove System Design and Simulation for Patients With Brachial Plexus Injuries," ser. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, vol. Volume 8: 47th Mechanisms and Robotics Conference (MR), 08 2023, p. V008T08A070. [Online]. Available: https://doi.org/10.1115/DETC2023-114907

[20] Y. Guo, W. Xu, S. Pradhan, C. Bravo, and P. Ben-Tzvi, "Integrated and configurable voice activation and speaker verification system for a robotic exoskeleton glove," in *International Design Engineering Technical Conferences and Computers and Information in Engineering*

Conference, vol. 83990. American Society of Mechanical Engineers, 2020.

[21] A. Rangapur, S. C. Sethuraman *et al.*, "Efficientword-net: An open source hotword detection engine based on one-shot learning," *arXiv preprint arXiv:2111.00379*, 2021.

[22] M. Wu, S. Panchapagesan, M. Sun, J. Gu, R. Thomas, S. N. P. Vitaladevuni, B. Hoffmeister, and A. Mandal, "Monophone-based background modeling for two-stage on-device wake word detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5494–5498.

[23] K. Kumatani, S. Panchapagesan, M. Wu, M. Kim, N. Strom, G. Tiwari, and A. Mandai, "Direct modeling of raw audio with dnns for wake word detection," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 252–257.

[24] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *2006 IEEE international conference on acoustics speech and signal processing proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[25] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5115–5119.

[26] Y. Zhang, M. Yu, N. Li, C. Yu, J. Cui, and D. Yu, "Seq2seq attentional siamese neural networks for text-dependent speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6131–6135.

[27] S. Kadam and V. Vaidya, "Review and analysis of zero, one and few shot learning approaches," in *International Conference on Intelligent Systems Design and Applications*. Springer, 2018, pp. 100–112.

[28] S. H. Mun, W. H. Kang, M. H. Han, and N. S. Kim, "Robust text-dependent speaker verification via character-level information preservation for the sdsv challenge 2020," *arXiv preprint arXiv:2010.11408*, 2020.

[29] A. Lozano-Diez, A. Silnova, B. Pulugundla, J. Rohdin, K. Veselỳ, L. Burget, O. Plchot, O. Glembek, O. Novotnỳ, and P. Matejka, "But text-dependent speaker verification system for sdsv challenge 2020." in *INTERSPEECH*, 2020, pp. 761–765.

[30] S. He, A. Zhang, and M. Yan, "Voice and motion-based control system: Proof-of-concept implementation on robotics via internet-of-things technologies," *ACMSE 2019 - Proceedings of the 2019 ACM Southeast Conference*, pp. 102–108, 2019.

[31] I. M. El-emary, M. Fezari, and H. Attoui, "Hidden Markov model/Gaussian mixture models (HMM/GMM) based voice command system: A way to improve the control of remotely operated robot arm TR45," *Scientific Research and Essays*, vol. 6, no. 2, pp. 341–350, 2011.

[32] J. B. Ǵomez, A. Ceballos, F. Prieto, and T. Redarce, "Mouth gesture and voice command based robot command interface," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 333–338, 2009.

[33] R. K. Megalingam, R. S. Reddy, Y. Jahnavi, and M. Motheram, "Ros based control of robot using voice recognition," in *2019 Third International Conference on Inventive Systems and Control (ICISC)*, 2019, pp. 501–507.

[34] M. Pleva, J. Juhar, S. Ondas, C. R. Hudson, C. L. Bethel, and D. W. Carruth, "Novice user experiences with a voice-enabled human-robot interaction tool," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, 2019, pp. 1–5.

[35] S. Guo, Z. Wang, J. Guo, Q. Fu, and N. Li, "Design of the speech control system for a upper limb rehabilitation robot based on wavelet de-noising," in *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*, 2018, pp. 2300–2305.