Mechatronics 83 (2022) 102745

Contents lists available at ScienceDirect

Mechatronics

journal homepage: www.elsevier.com/locate/mechatronics

Personalized voice activated grasping system for a robotic exoskeleton glove

Yunfei Guo [a], Wenda Xu [b], Sarthak Pradhan [b], Cesar Bravo [c], Pinhas Ben-Tzvi [a,b,*]

[a] Department of Electrical and Computer Engineering, Virginia Tech, United States of America
[b] Department of Mechanical Engineering, Virginia Tech, United States of America
[c] Carilion Clinic Institute of Orthopaedics and Neurosciences, Virginia Tech Carilion School of Medicine, United States of America

ARTICLE INFO

Keywords:
Natural language processing
Voice command system
Exoskeleton glove
SEA
Electronics design

ABSTRACT

This paper proposes a novel human machine interface (HMI) and electronics system design to control a rehabilitation robotic exoskeleton glove. Such system can be activated with the user's voice, take voice commands as input, recognize the command and perform biometric authentication in real-time with limited computing power, and execute the command on the exoskeleton. The electronics design is a stand-alone plug-and-play modulated design independent of the exoskeleton design. This personalized voice activated grasping system achieves better wearability, lower latency, and improved security than any existing exoskeleton glove control system.

1. Introduction

According to statistical data published in 2010, over 6.7 million of U.S. adults have difficulty grasping or handling small objects [1]. To potentially better the lives of such a large group of people, a robotic exoskeleton glove was designed to be used as a rehabilitation device for Activities of Daily Living (ADL) [2]. Many soft and rigid exoskeletons were proposed and designed by previous researchers.

The SEM glove designed by Nilsson et al. [3] has been used commercially. It consists of a soft glove powered by a motor with cable transmission. The SEM Glove is equipped with 3 FSR sensors placed at the middle and index finger pad and at the thumb. This glove can output a maximum of 5 N on one fingertip. However, a normal healthy 20 to 29 year-old male can output a maximum of 450 N on all fingers, which is about 90 N on each fingertip [4].

Ma et al. designed the SAFER rigid exoskeleton using cable transmission and strain gauges to perform force feedback [5]. This exoskeleton can output 10 N on each fingertip. The glove can be used for rehabilitation therapy [6,7]. Lee et al. built the iSAFER glove by adding a slip detection and motion amplification system on the SAFER glove design [8], which can be used as a rehabilitation device [9].

Refour et al. integrated an exoskeleton glove with series elastic actuators (SEA) [1,10]. Compared with the soft SEA glove, this glove can output 20 N on each fingertip. Chauhan et al. proposed a grasp prediction algorithm to enhance the previous motion amplification system [11,12]. Vanteddu el al, improved the structure of the glove [2] and added deformation control for more stable grasping [13].

Xu et al. designed an exoskeleton glove using a rigid and articulated linkage mechanism connected with a linear SEA on each finger, to control each of the linkage mechanisms [14]. A rotatory SEA is used to control the MCP joint of the thumb [15], and a linear SEA to control the wrist joint. This design is more compact than Refour et al.'s design and can output the same amount of force.

Exoskeleton gloves have been built and improved by many researchers; however, many challenges remain with respect to the control system design of an exoskeleton glove.

Li et al. proposed an EEG based method to control exoskeleton gloves with a 95.57% actuation success rate [16]. However, this system required a 8.06 s processing time before the actuation started.

Randazzo et al. proposed another EEG approach [17,18]. This method has less than 500 ms processing time. However, the earlier-mentioned system has a classification accuracy of around 70%.

Chowdhury et al. improved the grasp prediction accuracy to 75% by combining EEG with EMG signals [19].

Baklouti et al. proposed a vision-based system using head and mouth gestures [20]. This approach has trouble distinguishing between normal head and mouth movement and command gestures.

Kim et al. designed a vision-based system where users wear a camera for detecting the objects to be grasped [21]. This approach is only accurate when the user has a clear view of the target object with no overlap and no other objects in the view.

Researchers also proposed a number of voice-controlled exoskeletons [22–24]. These voice-based solutions lack biometric authentication and configurable activation keywords.

* Corresponding author at: Department of Mechanical Engineering, Virginia Tech, United States of America.
E-mail addresses: yunfei96@vt.edu (Y. Guo), wenda@vt.edu (W. Xu), sp21@vt.edu (S. Pradhan), cjbravo@carilionclinic.org (C. Bravo), bentzvi@vt.edu (P. Ben-Tzvi).

https://doi.org/10.1016/j.mechatronics.2022.102745
Received 8 December 2020; Received in revised form 11 November 2021; Accepted 16 January 2022
Available online 3 February 2022

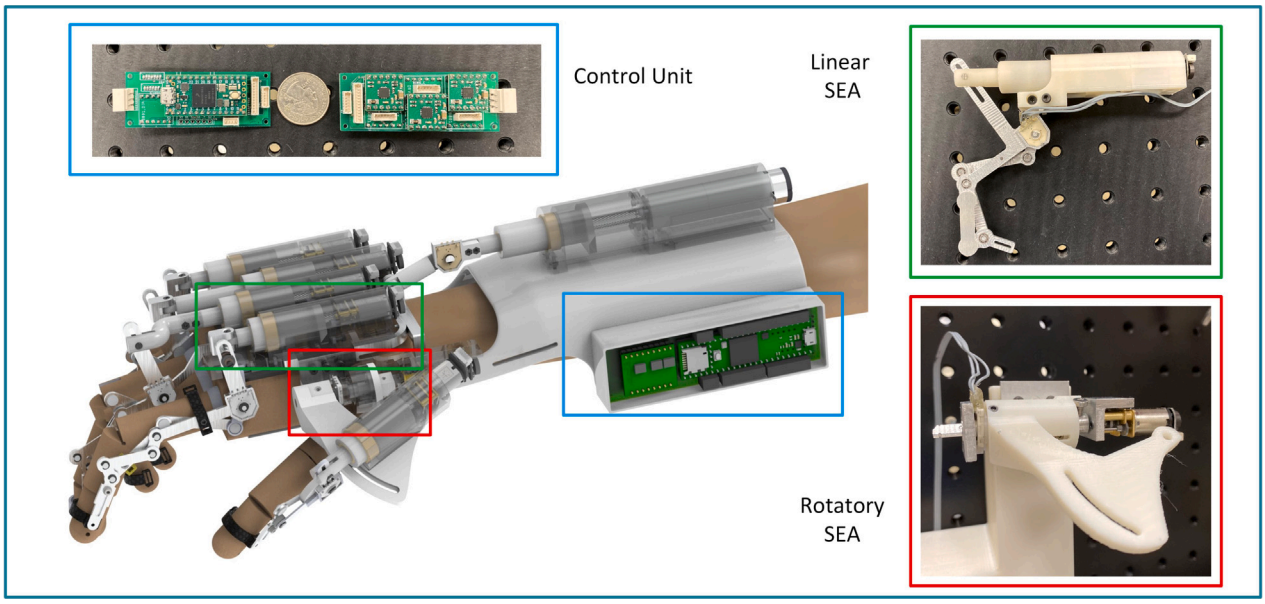

Fig. 1. Novel design of the new RML glove with seas and electronics.

This paper proposes a complete system to control exoskeleton gloves with portable hardware, slip detection algorithms, and voice-based HMI with bio-authentication feature. This grasping system includes a stand-alone plug-and-play modulated design electronics to control the exoskeleton, and a voice-control-based HMI called the integrated trigger-word configurable voice activation and speaker verification (CVASV) HMI. The electronics support slip detection algorithms, cloud computing, and provide over two hours of continuous usage. The CVASV HMI can perform customized voice activation and text-independent speaker verification on embedded systems with limited computing power. The following sections focused on the design of the electronics and CVASV HMI to achieve a stable grasp on an exoskeleton glove triggered by voice commands with bio-authentication.

2. Hardware introduction

In this paper, the personalized voice-activated grasping system is attached to Xu et al.'s glove design. However, this system's design principle is based on its ability to be easily deployed on any soft or rigid exoskeleton with force feedback sensors.

A Linear SEA and a rotatory SEA are shown in Fig. 1. Linear SEAs and articulated linkage mechanisms are used for all fingers on the RML glove. A linear SEA is used to perform indirect sensing of contact forces between the fingers and the grasped object. In Fig. 2, the red enclosed box points to an angular potentiometer which can measure the angle of the linkage attached to and actuated by the SEA. The distance denoted by AC can be calculated based on the angle measurement. The blue enclosed box points to a magnetic encoder that can measure the distance between points C and E. There exists a spring between points A and D, and CD and CB are of known value. The difference between AE and DE calculates the compressed spring length. The spring length calculation is shown in Eq. (1). Force can be calculated using Eq. (2).

$$L_{spring} = \cos(\angle ABC) \times CB + CE - DC - DE \tag{1}$$

$$force = k_{spring} \times \delta L_{spring} \tag{2}$$

The rotatory SEA is designed to duplicate the motion of the metacarpophalangeal (MCP) joint. The Rotatory SEA uses a similar structure and uses a torsion spring instead of a coil spring for the linear SEA. Fig. 3 shows the design of a rotatory SEA. The red enclosed box highlights the output shaft, and the blue enclosed box highlights the input shaft. There exists a torsion spring between the output shaft and the input shaft. The force can be easily calculated by measuring the difference between the input shaft and output shaft angles.

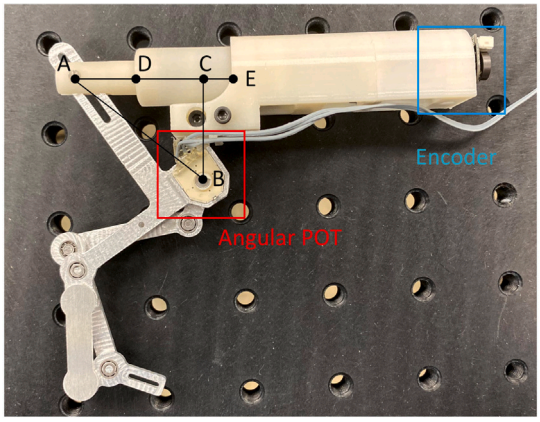

Fig. 2. Linear SEA installed on RML glove finger linkage mechanism. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Related work

3.1. Intelligent object grasping and learning system

Brielle [25] proposed an intelligent object grasping and learning system. The electronics design contains a WiFi model and Teensy 3.2 micro-controller with a stationary power supply. The size of the electronics is 5.3 cm × 8 cm × 4 cm, which is relatively large for exoskeleton gloves and does not have good wearability. The mobility of the patient's arm will be dramatically affected if a 5.3 cm × 8 cm × 4 cm controller box is placed on the patient's hand or arm. If the controller box is placed on the waist, it requires extra wiring from the waist to the exoskeleton, which will also affect the comfort of wearing. The low-level control programs run in a state machine using a single thread which bottlenecks sensor readings. Therefore, the system experiences noticeable delays while controlling all fingers simultaneously.

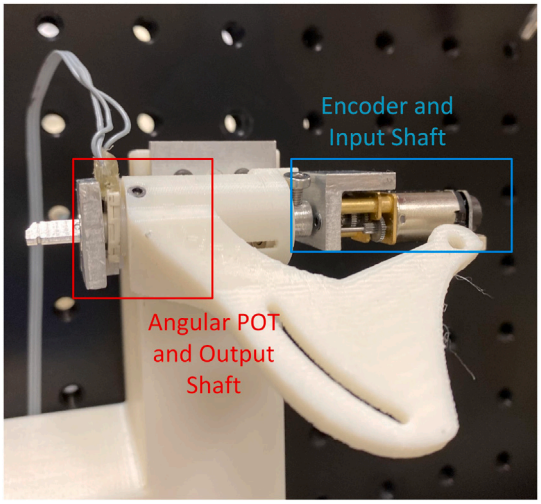

Fig. 3. Design of the rotatory SEA on the RML glove thumb linkage. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Force sensors are placed on the fingertips to measure and regulate contact forces. The force sensors are also used to monitor any force changes on the fingertips. When a grasped object starts to slip, the force measured on the fingertips would inherently decrease dramatically. Therefore, if there is a dramatic decrease in the force-sensor reading, it indicates that the object grasped by the user is slipping. The HMI system uses the user's hand movement to initiate a grasp—the system is activated by sensing a minor twitch of the finger. Whenever the force sensing on the finger changes, a grasp will be initiated.

This system requires the user to initiate a grasp with finger movement. However, not all patients suffering from hand disability have the ability to control their fingers. This method is only suitable for patients with some finger movement capabilities and cannot be used for patients who cannot move or control their fingers.

3.2. VGG-M speaker verification

VGG-M is a deep learning approach to the voice verification system proposed by Nagrani et al. [26]. The VGG-M speaker verification method is used as the performance baseline method in this research. This deep-learning voice verification system achieves a better Equal Error Rate (EER) (10.2%) than the non-deep learning baseline (15.0%).

During training, each audio file is divided into several 3-second audio clips. Each audio clip is turned into a 512 × 300 spectrum using Fourier transform. The spectrum is treated as 2D images and fed into an utterance-level feature extractor. With the utterance-level features, 1251 speakers from VoxCeleb1 are classified as 1251 classes by the VGG-M deep neural network.

During testing, the VGG-M network is used as a feature extractor. VGG-M changes the average pooling layer to match the test-time length so that the network can take inputs of various lengths. If the cosine of the distance between two different audio samples is within a threshold, the two samples are considered to be of the same class. The paper also proposes a test-time augmentation method which randomly selects ten samples from the entire data set and calculates the average distance between features. The VGG-M network with softmax loss, global average pooling, and test time augmentation 2 was used as the deep learning baseline in this research.

3.3. MobileNet

When using personalized voice activation and command system on exoskeletons, computation speed is crucial. The networks proposed by Nagrani et al. [26] are not the fastest networks to run on a mobile device. As such, Andrew et al. [27] proposed an efficient convolution neural network used on image classification. MobileNet uses a deep-wise separable convolution to replace the traditional convolution. $C_{traditional}$: Computation cost of traditional convolution. The cost of a traditional convolution is shown in Eq. (3). The deep-wise separable convolution is shown in Eq. (4). Compared to traditional convolution, the MobileNet's deep-wise separable convolution is faster based on Eq. (5).

$$C_{traditional} = H_i \times W_i \times C_i \times H_k \times W_k \times C_o \tag{3}$$

$$C_{mobilenet} = H_i \times W_i \times C_i \times H_k \times W_k + C_i \times C_o \times H_i \times W_i \times W_k \tag{4}$$

$$\frac{C_{mobilenet}}{C_{traditional}} = \frac{1}{C_o} + \frac{1}{H_k \times W_k} \tag{5}$$

where $C_{mobilenet}$ is the computation cost of MobileNetV1, H_i is the height of the input array, W_i is the width of the input array, C_i is the channel of the input array, H_k is the height of the kernel, W_k is the width of the kernel, and C_o is the channel of the output array.

MobileNet-224 (MobileNetV1) achieves a similar classification accuracy (70.6%) as VGG16 (71.5%) on the ImageNet dataset. MobileNet-224 has far fewer parameters (4.2 million) than VGG16 (138 million); thus, it is faster than VGG16. The VGG-M network is modified based on VGG16, which has a similar accuracy in image classification. It is possible to use this method to accelerate the existing speaker verification process.

4. Proposed approach

The grasping system shown in Fig. 4 contains a smartphone, a microphone, and a micro-controller. The microphone inputs raw data into the smartphone, and all the personalized voice activation and verification are calculated onboard. The user will wear a microphone placed on the collar or shirt, and the microphone is connected by a wire to a smartphone located in the patient's pocket. The smartphone and microcontroller communicate through Bluetooth. The micro-controller is responsible for processing signals from sensors and perform force feedback control and slip detection.

Unlike Apple's trigger word "Hey Siri" or Google's trigger word "Ok Google", the trigger word proposed in this paper can be customized. Apple and Google use text-dependent voice activation that require users to activate the system using predefined trigger words. The system proposed in this paper uses text-independent voice activation, which means the user can define the trigger word they prefer. Also, Apple and Google only perform speaker verification on the trigger word after the system is activated. Several problems can be caused by using the built-in voice assistant. For example, when the user has dinner in a restaurant, the system is continuously activated due to frequent grasping of utensils while eating. If someone sitting next to the user said something similar for example to "release.", then the system may pick up that command and drop his utensil.

4.1. Electronics and low level control

The electronics design is portable, has low latency, and modularized. The electronics and power supply can be attached and detached easily. This design can be easily modified and attached to other exoskeleton gloves other than Xu et al.'s glove.

The micro-controller, motor-controllers, and power supply are in three separated units connected with MOLEX connectors as shown in Fig. 5. The battery in the power supply unit can be easily disconnected

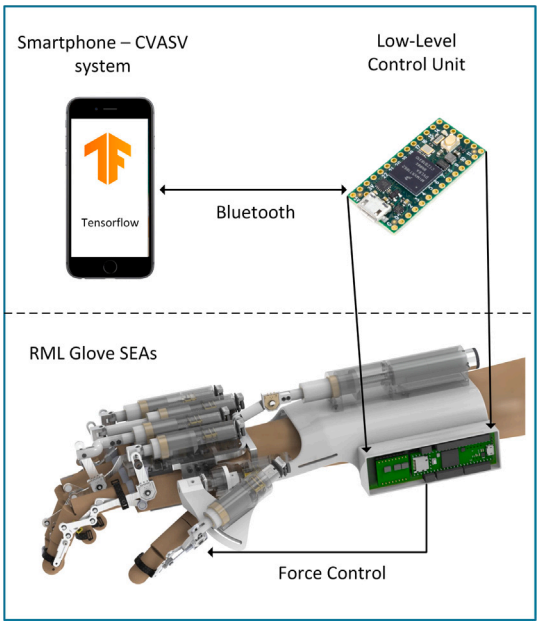

Fig. 4. Personalized voice activated grasping system overview.

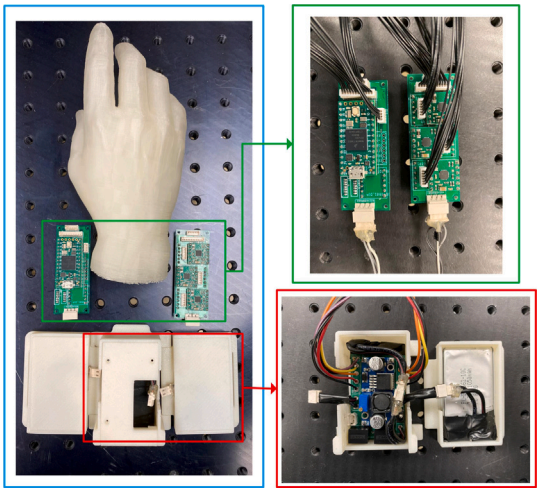

Fig. 5. Electronics: (Top right) Side mount computational unit and motor control unit. (Bottom right) Power conversion unit with battery.

and changed to a different battery with lower or higher capacity. The voltage converter unit dimensions are 55 mm × 35 mm × 15 mm, and the battery dimensions are 62 mm × 35 mm × 18 mm. The micro-controller and motor controllers are separated, and each has a size of 62 mm × 25 mm × 15 mm. The size of each unit is compact and can be easily mounted on both sides of the arm.

Three separate PCBs were designed to place all the components on the computational box and power conversion. The components overview of each layer is shown in Fig. 6.

The onboard microcontroller is responsible for sensor reading and performing low-level control, including force control and slip detection. The slip detection uses the same idea as Lee [25] proposed to measure the force change on the fingertip to detect slip. Instead of using a force sensor, the SEA is used as the force sensing device. An integrated wearable battery box will supply power for two hours of continuous operation.

Encoders and angular potentiometers are attached to measure each finger and wrist angle and calculate the force output of the SEAs. A

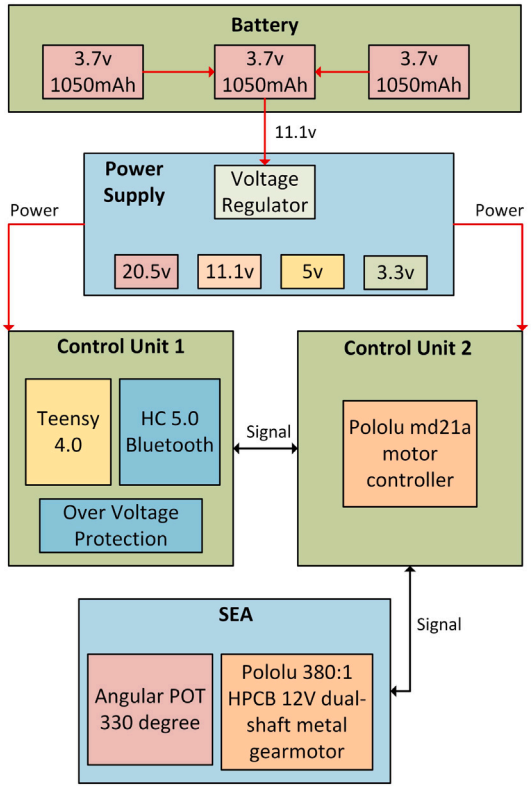

Fig. 6. Electronics overview.

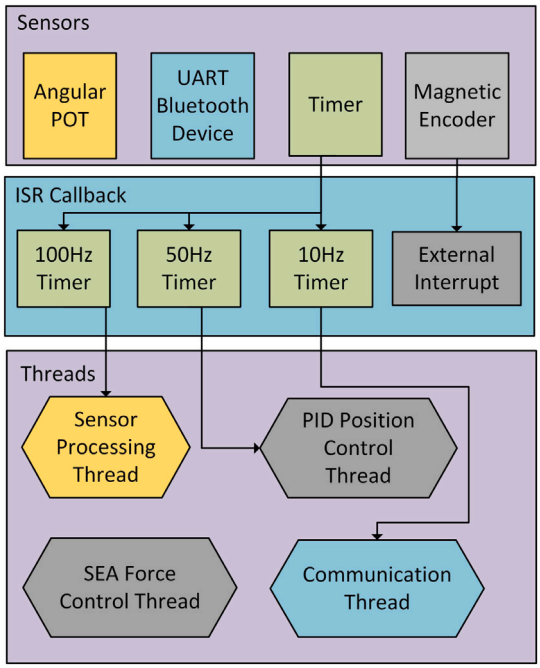

Fig. 7. Structure of low-level control.

Teensy 4.0 micro-controller was selected to read data from sensors and perform low-level control. A real-time system was used to minimize sensor reading latency and provide the ability to perform parallel computing. The structure of the low-level control system is shown in Fig. 7.

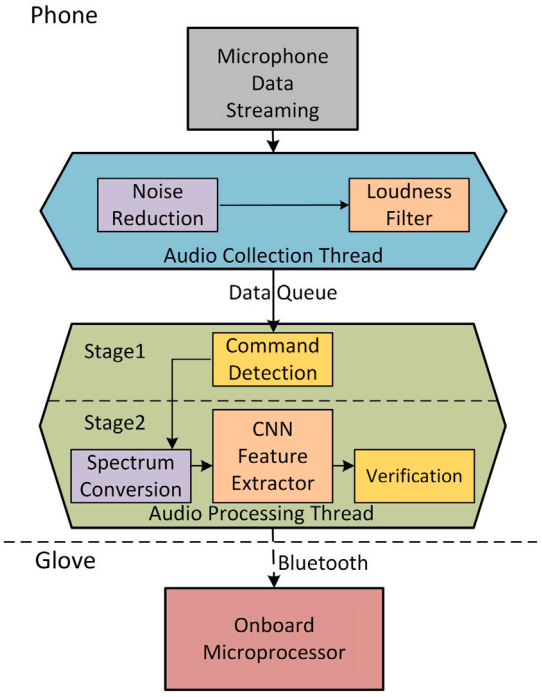

Fig. 8. Structure of the CVASV HMI.

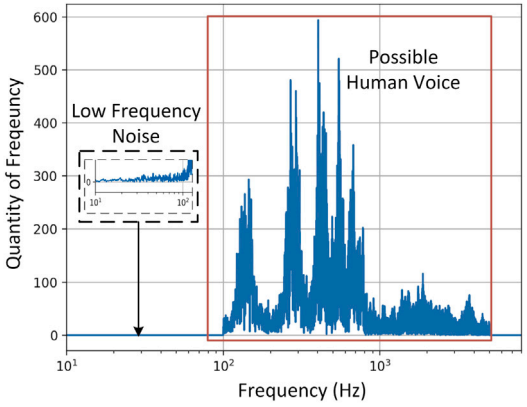

Fig. 9. Noise filter.

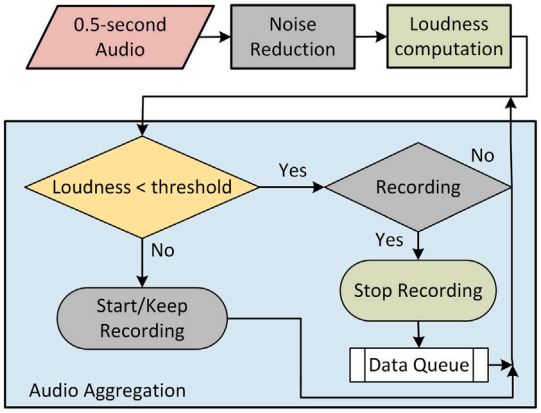

Fig. 10. Flowchart of loudness detection.

4.2. CVASV HMI

The configurable personalized voice activation and command system can be divided into two sections. The first section controls the triggered activation process. The loudness level of the input audio is used as a trigger. If the loudness is above a certain threshold, the system will pass the data into the verification section.

The verification section contains the verification process. The keyword is verified using the Google Web Speech API, first. If the keywords are correct, the data is passed into the verification model. If the verification result showed high similarity with the registered user, the command is accepted. Fig. 8 shows the thread assignments and the tasks that occur in each thread. The configurable voice activation section takes the raw audio data as input and outputs the accepted audio data. The microphone streaming callback continuously generates 0.5-second audio segments. The segments are processed by the noise reduction and loudness filter in the audio collection thread. Commands are likely to be spread in between multiple audio segments. The audio collection thread's job is to combine these audio segments into commands based on the loudness. For example, if a "hey glove" command is located in 3 separate audio segments, the audio collection thread will combine these three segments into one 1.5 s duration audio segment.

The audio collection thread contains a noise reduction filter and a loudness filter. If the loudness is greater than the threshold and a complete sentence has been detected, the audio collection thread enters the pre-active mode. The audio data queue then sends data from the audio collection thread to the audio processing thread under pre-active mode. The voice processing thread consists of two stages. The command detection stage uses voice recognition API. After the command is accepted, it enters the speaker verification stage. The MobileNetV1 speaker verification system verifies if the audio belongs to one of the enrolled speakers.

4.3. Configurable voice activation

The voice activation system includes a noise reduction filter, a loudness filter, and a command detector. The system is designed to detect possible activation commands using minimal computational cost. The system enters a pre-activated mode if the user's voice intensity is greater than a threshold. A human-like voice is checked using voice activation API to verify if an activation trigger-word is present. The system then enters into an activated mode, and the audio is passed to the voice verification system.

The raw input of the microphone voice contains high and low-frequency noise. With a noise reduction filter, the system will enter the pre-active mode less frequently from false-positive activations in a noisy environment.

A noise-reduction filter is used to minimize high and low-frequency noise. It converts raw input from the microphone to the time-independent frequency domain using a Fourier transform. The total relative power density within each frequency band is measured. High-frequency and low-frequency bins are considered to be noise and are subsequently trimmed from the spectrum. The time-independent Fourier transform [28] of the original audio data and the filtered audio data are shown in Fig. 9. The remaining frequencies contain potential human voices, which then enter the loudness detector.

The loudness detector aggregates small audio segments into complete sentences. Audio data containing a loudness level above a certain threshold sets the system into pre-active mode. The input data is a 0.5-second filtered audio segment. These audio segments are grouped into a larger audio segment based on the margin between each word. If the margin between words is longer than 0.5-seconds, the sentence ends. The loudness detector with audio aggregation is implemented according to the flowchart described in Fig. 10.

After the loudness detector detects a complete sentence, the audio data is sent to the audio processing thread for voice recognition. The

loudness detector keeps monitoring the environment. In the voice processing thread, the complete sentence will be recognized using the Google Web Speech API. However, any voice recognition API can be used. Voice recognition APIs from Amazon, Google, or Apple are proven to be accurate and require a low computational cost.

If the system is not activated, the API result is used for activating the trigger-word detection. The API result is forwarded directly as an unverified command if the system is already activated. Words with similar pronunciation to the activation trigger-word or the command itself are accepted. After checking that the trigger-word or command is valid, speaker verification is ultimately performed using the filtered audio data.

4.4. Voice verification

The voice verification method was improved based upon Nagrani et al. [26] VGG-M with the Softmax loss function speaker verification method. Instead of using VGG-M, the low computational cost network MobileNetV1 was used. The size of the MobileNetV1 network was modified by reducing it to 75% of the original network such that the alpha value equals 0.75. The network was modified to adapt 2D spectrum input.

The training procedure for MobileNetV1 is similar to the baseline VGG-M method. Filtered audio data from the previous speaker activation section is converted into a 2D spectrum [28] with both frequency and time information. The 2D spectrum is fed into MobileNetV1. The speaker utterance is grouped using the Global Average Pooling (TAP) into 768 features and then classified using a dense layer. The loss function used for classification was the standard Softmax loss function given by:

$$P(y=j|x)=\frac{e^{X^T W_j}}{\sum_{k=1}^{K} e^{X^T W_k}} \tag{6}$$

where x is the spectrum input, j is the label of a certain speaker, X is the output of the feature from the neural network, k is the number of classes, and w is the weighting vector. The training process is shown in Fig. 11.

During the speaker verification procedure, the last dense layer of the MobileNetV1 is detached. The output of the network has 768 features. The distance between the two audio features vectors A and B is calculated. If the distance is less than an enrollment threshold, these two audios are from the same speaker. The verification procedure is shown in Fig. 11. The distance between features is calculated by the cosine of the distance [29] (D_c) defined as:

$$D_c=\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{7}$$

During the enrollment process, the user needs to record a x-second long enrollment audio. The audio is then randomly cut into n audio segments, each having a 3-second length. These n audio clips are processed by the neural network, and each audio clip is turned into a feature vector (F). The average enroll-feature vector (F_{avg_enroll}) can be calculated through Eq. (8).

$$F_{avg}=\frac{\sum_{k=1}^{n} F_n}{n} \tag{8}$$

During the verification process, the input test audio is cut into n audio segments, each having a 3-second length. The average test-feature vector (F_{avg_test}) can also be calculated through Eq. (8).

The average distance between the same speaker and different speakers over a verification training data set can be calculated and used as the verification threshold. The verification training data set contains K speakers. The threshold (T) is chosen by calculating the mean distance between the average enroll-feature (F_{avg_enroll}) and the average test-feature (F_{avg_test}) over K speakers, as shown in Eq. (9).

$$T=\frac{\sum_{k=1}^{K} dist(F_{avg\_enroll}, F_{avg\_test})}{K} \tag{9}$$

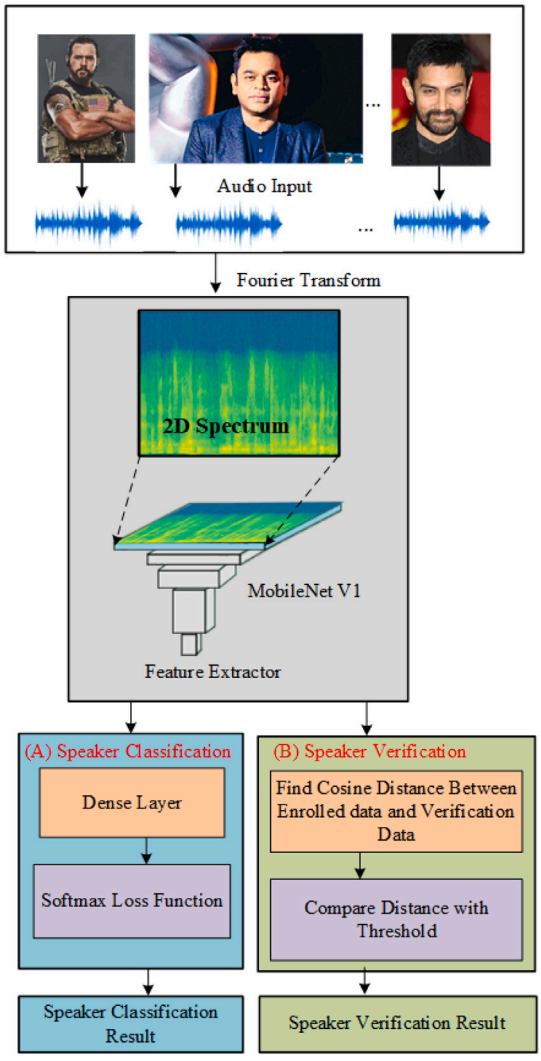

Fig. 11. Training and testing with modified MobileNetV1. (A) Training add-on. (B) Testing add-on.

Table 1
Latency in low level control of the personalized voice activated grasping system.

	Designed time cost	Latency
Sensor reading	10 ms	0 ms
SEA position control	10 ms	0 ms
SEA force control	110 ms	0 ms
Bluetooth communication	100 ms	2 ms

5. Experiments

5.1. Slip detection and low-level control

The latency of the low-level control has been tested. Three major components can cause latency: sensor reading, SEA position control, and SEA force control. The micro-controller used is a Teensy 4.0 @ 800 MHz, and latency was measured for all three components. The results are provided in Table 1. The experiment demonstrated that the multi-threading system is well designed, where the low-level control only has a total of 2 ms latency caused by the Bluetooth communication.

According to Lee et al. [25], detecting the force differences between different fingertips can indicate slippage of the grasped object. In this

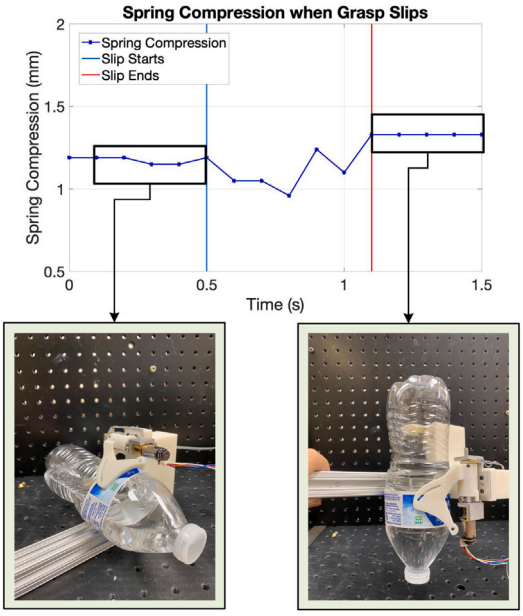

Fig. 12. Slip detection using linear SEA.

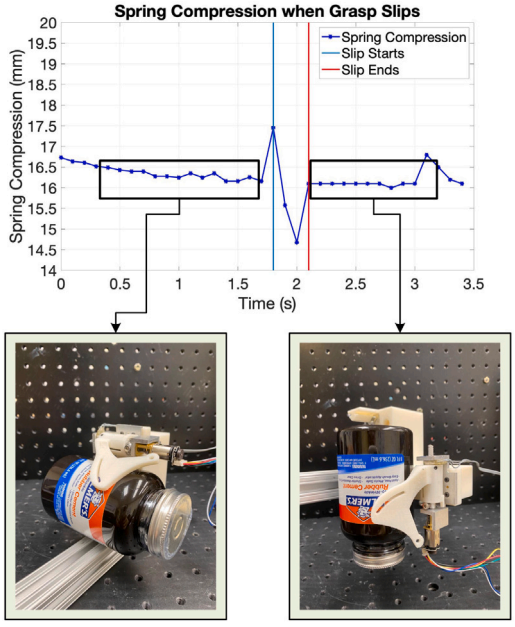

Fig. 13. Slip detection using rotatory SEA.

experiment, instead of using a resistor-based force-sensor, an SEA was used to measure contact forces. One experiment demonstrated that a plastic bottle could be stably grasped with minimal force using a single finger SEA. The single finger SEA is manually rotated by 90 degrees and shaken until slip can be visually observed. During this process, spring compression vs. time is recorded. The results of the linear SEA readings are shown in Fig. 12, and the results of the rotatory SEA readings are shown in Fig. 13. Spring compression is proportional to force; thus, the change in spring compression is proportional to the change in force. The SEA detects the slip, and the low-level force feedback control can cause the SEA to apply the appropriate level of force to eliminate slip.

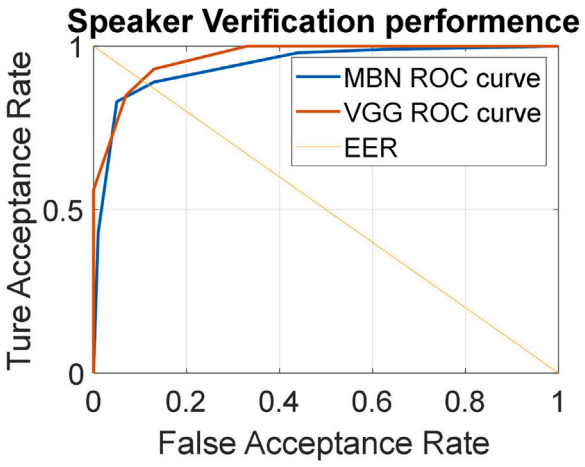

Fig. 14. ROC curve of CVASV verification on VoxCeleb1 test dataset.

5.2. CVASV HMI

5.2.1. Feature extractor

The feature extractor training is performed on the VoxCeleb1 dataset. The VGG-M method will be used as a performance baseline. The VoxCeleb1 [30] dataset was used for training the voice verification system in this research. This dataset contains 1251 celebrities giving presentations in different environments. There are over 100,000 utterances, and 40 speakers are chosen for the verification test. The dataset contains noisy data, which is suitable for training a robust speaker verification model. A 3-second audio clip is randomly extracted from each audio file and converted to a 512 × 300 spectrum using Fourier transform. The 512 × 300 spectrum is used to train both MobileNetV1 and VGG-M baseline networks. All networks will classify input audio data into 1251 classes. The classification result is shown in Table 2.

5.2.2. Speaker verification

The speaker verification dataset is from the VoxCeleb1 verification dataset. It contains 40 speakers that have IDs from 270–309. Two audios are given during the verification testing process: one is used as enrolled data, and the other is used as test data. The verification test data and the enrolled data are randomly divided into three segments, consisting of 3-second audio. The time cost to process one 3-second audio clip will be measured to compare the performance. The result can be visualized by the receiver operating characteristic (ROC) curve shown in Fig. 14.

The performance of using MobileNetV1 is now compared to the original method, which uses VGG-M. The performance comparison between the modified MobileNetV1 (MBN) and the VGG-M is shown in Table 2. The modified MobileNetV1 achieves a 0.8% higher accuracy in classification. The modified MobileNetV1 has a 2.4% higher ERR rate than the baseline VGG-M in the verification section. However, using MobileNetV1 as the feature extractor reduces the total parameters in the neural network by 89.6% and was on average 17.7% faster than using VGG-M.

Table 3 compares the verification method with state-of-the-art speaker verification methods on the Voxceleb1 dataset. The baseline method in this paper is slightly different than the ERR rate as Nagrani, et al. [26] proposed due to the different aggregation method. The test data in VoxCeleb1 is longer than 3 s and has a variable length. We chose not to use variable average pooling to adapt different input lengths due to the fact that our application is used for short commands that are less than 3 s long.

Table 2
HlComparison between Modified MobileNetV1 and the baseline VGG-M method.

	MBN	VGG-M
Classification accuracy	**80.0%**	79.2%
ERR rate	12.4%	**10.0%**
Time cost	**73.2 ms**	88.9 ms
Total parameters	**1,832,544**	17,691,328

Bold text highlights better performance.

Table 3
Comparison between the Modified MobileNetV1 and other state-of-the-art speaker verification methods on VoxCeleb data set.

Author	Network	Aggregation	EER
Proposed	MBN	Rand3S	12.4%
Proposed	VGG-M	Rand3S	10.0%
Nagrani et al. [26]	GMM-UBM	N/A	15.0%
Nagrani et al. [26]	I-vectors+PLDA	N/A	8.8%
Nagrani et al. [26]	VGG-M	TAP	10.2%
Chung et al. [31]	VGG-M-SC	TAP	5.94%
Xie et al. [32]	Thin-ResNet-34	TAP	10.48%
Xie et al. [32]	Thin-ResNet-34	GhostVLAD	3.22%

TAP: Temporal Average Pooling.
VGG-M-SC: VGG-M network with softmax and contrastive loss.
Rand3S: average of random 3-second audio clip.

Table 4
CVASV HMI performance test at cosine distance threshold = 0.25.

Speaker	TA rate	FA rate
A	95%	3.1%
B	100%	9.3%
C	100%	10.7%
D	100%	8.6%
E	100%	12.4%
F	92%	4.9%
G	92%	4.4%
H	100%	11.2%
I	96%	7.6%
J	96%	6.2%

Table 5
Time cost to process a one audio command.

	Time cost
Noise reduction filter	2.12 ms
Loudness filter (D_e)	2.45 ms
Audio collection subsystem	5.02 ms
Spectrum conversion	12.32 ms
Voice recognition API	34.47 ms
MobileNetV1 feature extractor	126.2 ms
Voice recognition subsystem	182.58 ms

5.2.3. RML exoskeleton voice control dataset

This dataset contains ten-speaker models that contain six males and four females, including Asian accents, Indian accents, and native North American English speakers. Each speaker model is made up of 2 sections: an enrollment section and a testing section. The enrollment section contains 5 commands as enrollment data: “hey glove”, “grasp bottle”, “grasp cup”, “grasp toothbrush”, and “release”. The enrollment data is recorded when each speaker reads these commands in a quiet room with normal speed and tone. This section is used to enroll the speaker and tune the CVASV HMI. The testing section contains 20–25 short commands for each speaker, recorded with various background noise. This section is used to test the performance of the CVASV HMI. This dataset can be used on other voice command systems with similar design principles as well.

5.2.4. CVASV HMI performance on RML dataset

The CVASV HMI is configured to use Google Web Speech API for commands classification. The text generated by Google API is classified into 6 categories by the number of matching characters, including 1 activation command, 4 grasp commands, and an unknown command. The classification performance is measured by the classification accuracy shown in Eq. (10).

$$Acc = \frac{N_{correct}}{N} \tag{10}$$

Acc – Classification accuracy

$N_{correct}$ – Number of correctly classified commands

N – Number of classified commands

The verification section uses MobileNetv1 as a feature extraction network and uses a cosine distance to measure speakers’ similarity. After observing the ROC curve, the threshold is set to 0.25 to achieve a good true acceptance (TA) rate while avoiding false acceptance (FA).

The first experiment was conducted to test the performance of the CVASV HMI on RML dataset. During the enrollment, each speaker has five enrolled models which represent each different command. The enrollment model in the RML dataset is used to enroll each speaker. All the commands are tested in two parts: classification and verification. The performance of classification is evaluated by the Command Acceptance rate shown in Eq. (10), which resulted in a classification accuracy of 94.1%.

The accepted command is then verified using the verification model. CVASV HMI’s performance is quantified using the true acceptance (TA) rate and false acceptance (FA). The verification true acceptance (TA) rate is defined by Eq. (11). The higher the TA rate, the better the model is. The verification false acceptance (FA) rate is defined by Eq. (12). The lower the FA rate, the better the model is. The results are shown in Table 4.

$$TA_{Rate} = \frac{N_{TA}}{N_{same}} \tag{11}$$

$$FA_{Rate} = \frac{N_{FA}}{N_{diff}} \tag{12}$$

TA_{Rate} – True acceptance rate

N_{TA} – Number of commands verified to be from the same user are from same users

N_{FA} – Number of commands verified to be from the same user are from different users

N_{same} – Number of commands from the same user

N_{diff} – Number of commands from different users.

The second experiment tested the latency of the system. The CVASV HMI is designed to run in real-time on a portable device reliably. This experiment was done using a computer with a 2.2 GHz six-core Intel i7 processor to simulate the mobile device. The program uses only one thread with the CPU frequency limited to 2 GHz and RAM limited to 2 Gigabytes. The latency for each component is shown in Table 5. The entire system’s latency is 182.58 ms, which is faster than most HMIs and can run in real-time on a portable device.

The third experiment was conducted to study the impact of noisy data on CVASV HMI. The same command of the same speaker under different background noise was compared. The results are shown in Table 6. The verification model has higher cosine distance under noisy environment.

The fourth experiment was conducted to study how biological gender influences the verification system. Speaker A and speaker B are of opposite sex. Speaker A is enrolled and tested with speaker B’s testing

Table 6
Cosine distance between same speaker under different background noise.

Command	Noisy dis	Quiet dis
"hey glove"	0.17	0.12
"grasp toothbrush"	0.13	0.11
"grasp bottle"	0.12	0.06
"grasp cup"	0.12	0.09
"release"	0.21	0.13

Table 7
Cosine distance between different sexes.

Command	Same sex dis	Oppos. sex dis
"hey glove"	0.52	0.48
"grasp bottle"	0.37	0.39
"grasp cup"	0.47	0.42
"grasp toothbrush"	0.44	0.39
"release"	0.28	0.51

Table 8
Cosine distance of speaker with different accents.

Command	Distance
"hey glove"	0.32
"grasp toothbrush"	0.34
"grasp bottle"	0.42
"grasp cup"	0.38
"release"	0.47

data. Theoretically, none of these commands will be accepted. Speaker A is then enrolled and tested with speaker C's testing data, where Speakers A and C are of the same sex. Theoretically, none of these commands should be accepted either. The results of this experiment are shown in Table 7.

There were no significant differences between the two scenarios when performing verification over speakers of the same and opposite sex. All the test data performed above the 0.25 distance threshold as expected.

The fifth experiment was done to study how the accent will affect the CVASV HMI. Speaker E and Speaker C are of the same gender. Speaker E is a native English speaker without a noticeable accent. Speaker C is a fluent English speaker with a noticeable accent. From Table 4, a speaker with an accent will have lower accuracy in the recognition system, but the verification system is not affected. In this experiment, verification accuracy was tested. Speaker E is enrolled and tested against speaker B's testing data. Theoretically, none of these commands will be accepted. The results are shown in Table 8. The results proved that accent does not affect the verification system. All distance values are higher than the 0.25 threshold as expected.

The sixth experiment tested whether speaker verification can verify the same speaker with a genuine verification model. As mentioned previously, for each speaker, each command has its enrollment model. This experiment was performed to verify whether using a genuine model for all commands is feasible. All the enrollment commands are combined and enrolled as a genuine model for speaker A, and all test data was tested against this command. The results are shown in Table 9. The same speaker saying different commands has a similar physical distance to different speakers saying the same command. The commands are too short to distinguish between similar voices. To use a genuine model to perform speaker verification, the command needs to be made longer. However, using a longer command is not practical. The current solution is to have a separate model for each speaker containing every command. Using a separate model for each command and speaker will allow the speaker verification network to distinguish between different speakers.

Table 9
Cosine distance between different commands of same speaker.

Command	distance
"hey glove"	0.34
"grasp toothbrush"	0.42
"grasp bottle"	0.33
"grasp cup"	0.33
"release"	0.33

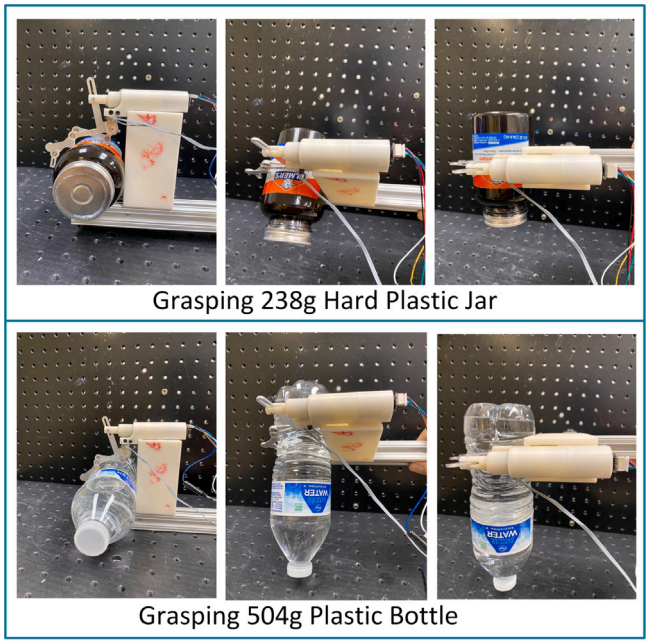

Fig. 15. Grasping with personalized voice activated grasping system using linear SEA.

5.3. Personalized voice activated grasping system

on the SEAs. Both rotatory and linear SEAs are used in the experiment. This experiment is conducted to prove the feasibility of the system. The experiment with hardware is to demonstrate that all the subsystems can work together with no problem. The high-level control program runs on a computer with limited RAM and CPU thread number to simulate a cell phone. The microphone used is the integrated microphone on a wired Apple Ear-pod connected to the simulated cell phone. The user is wearing the ear-pod while speaking the command. The experiment includes using the CVASV HMI to initialize the grasp, using SEA to apply proper force, and using slip detection to maintain its stability. The results are shown in Figs. 15 and 16. The CVASV HMI successfully initiated the grasp. The electronics and low-level control functioned flawlessly to grasp a heavy object. When the angle changes, the slip detection algorithm can detect slip and apply additional force incrementally to maintain a stable grasp. The latency between inputting a voice command to the glove and initiating a grasp is 184.58 ms. This latency proved the CVASV HMI, and the low-level control can run in real-time with low latency.

5.4. Energy consumption

The energy consumption experiment performed measures the energy consumption of the control unit, the rotatory SEA, the linear SEA, and CVASV HMI. The power consumption is measured through the power supply current. The result shows the power consumption averaged within 5 s. The SEAs' power consumption is measured with the control unit connected. The result is shown in Table 10.

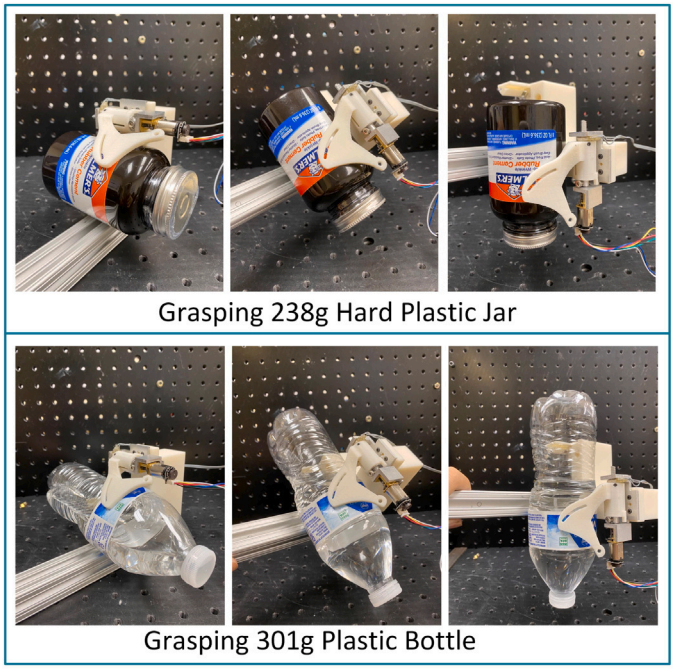

Fig. 16. Grasping with personalized voice activated grasping system using rotatory SEA.

Table 10
Power consumption of the exoskeleton.

Part	Status	AVG power (W)
Control Unit	idle	0.66
Control Unit	active	0.9
Rotatory SEA	idle	0.73
Rotatory SEA	hold grasping	1.37
Rotatory SEA	moving	2.68
Linear SEA	idle	0.72
Linear SEA	hold grasping	0.97
Linear SEA	moving	3.37

Table 11
Power consumption of CVASV HMI.

Status	Loss in battery	Power (W)
idle	9%	2.925
active	21%	6.825

The CVASV HMI has not been deployed onto a smartphone yet. The power consumption experiment is designed to approximate the battery usage on a smartphone and improve the system before it is deployed on a smartphone. A laptop with a 5953 mAh, 11.4 V battery is used in this experiment. The screen is turned off, and no additional programs were running other than the CVASV HMI. The CPU frequency is limited to 2 GHz, and the RAM is limited to 2 Gigabytes. The result is measured in loss of battery percentage. Idle status measured the power consumption when the system is placed in a noisy room for 2 h. Under these conditions, the CVASV system will constantly go into pre-active mode due to the above threshold noise. Active status measured the power consumption when the system is placed in the same noisy room. The system is activated 50 times and input the "grasp", "Release", "Stop" commands. The results are shown in Table 11.

6. Future work

6.1. Electronics and hardware

The experiments on a single finger proved that the electronics and hardware fulfill all the requirements. The entire glove needs to be manufactured and assembled to enable further testing of the complete system.

6.2. Command recognition subsystem

Command classification using Google Web Speech API has achieve 94.1% command classification rate, 97.1% ture acceptance rate and a 35 ms time cost. However, it requires an internet connection to perform web computing. Connection to the internet is crucial to the CVASV HMI, which makes it less portable. There might be situations where the patient does not have internet access. In the future, we will consider replacing the Google Web Speech API with local HMM based command classifiers, such as, LD3320 speech chip, Julius API, or PocketSphinx.

6.3. Speaker verification subsystem

The verification system achieves an average of a 97.1% true positive rate and a 7.84% false positive rate when tested with the RML voice dataset. However, there are still some improvements that can be made. Nagrani et al. [26] has proven that using large margin Softmax loss function [33] will improve the classification accuracy of VGG-M. Extending the training dataset to VoxCeleb2 [31] should improve the accuracy without any detrimental effects. The speaker verification system consumes a relatively large amount of energy. In applications where a stable network is available, we recommend using cloud computing for faster speed and low energy consumption.

7. Conclusion

The Personalized Voice Activated Grasping System proposed in this research was proven to be fast, accurate, portable, and a secure method to control Xu et al.'s exoskeleton glove through voice activation. The electronics and hardware could function without issues for more than 100 h of testing and proved to be reliable. The functionality was complete and was able to provide force feedback, force control, and connectivity. The electronics and hardware performed well and did not require any improvement. The low-level control system was fast and accurate, and the force feedback SEAs were accurate and could detect slip. The CVASV HMI can distinguish between different speakers and recognize different commands. The entire system has a less than 200 ms latency and has an average 91.4% chance to classify and verify the command correctly. Table 12 shows the performance comparison between the personalized voice activated grasping system and other state-of-the-art short voice command systems.

The design principle of the intelligent force grasping system is fully configurable. The application of this system is not limited to specific exoskeleton gloves. This paper uses the Xu et al.'s glove to better understand how the components are designed. The power supply unit can be quickly disconnected and replaced with any power supply unit based on different applications. The low-level electronics can be replaced by any compact design and can provide sufficient computing power to control the exoskeleton. The CVASV HMI can also be programmed to adapt different automatic speech recognition (ASR). This paper aims to provide voice recognition and speaker verification based on a fully customizable system that can be used and improved by other researchers.

The Personalized Voice Activated Grasping System will be further tested on the RML glove after the complete glove is finished. The RML exoskeleton dataset will be extended to 50 speakers with data collected

Table 12
Comparison between personalized voice activated grasping system and other state-of-the-art voice command system.

Author	Method	Acc*	CV-SV
Proposed	GoogleAPI+CNN	91.4%	Yes
He et al. [34]	GoogleAPI	92%	No
El-emary et al. [35]	GMM	< 85%	No
Gomez et al. [36]	MG GMM+SM	88%	No
Gomez et al. [36]	HMM	100%	No
Megalingam et al. [37]	PocketSphinx: HMM	90%	No
Pleva et al. [38]	Julius: HMM	91%	No
Guo et al. [39]	LD3320 speech chip	94%	No

CV-SV: Customized voice activation and speaker verification.
Acc*: For HMIs without speaker verification, Acc is the classification accuracy. For this paper, Acc stands for system accuracy, which is the classification accuracy times the true acceptance rate at 0.25 threshold.
GMM: Gaussian Mixture Model.
MG GMM+SM: Mouth gesture based detection using GMM and state machine.
HMM: Hidden Markov Model.

from real patients' clinical trials. With the data from actual patients, the system can be modified to assist patients.

The CVASV HMI will be applied on a smartphone and tested with a larger dataset. Suppose the recognition accuracy is not satisfactory during further testing. In that case, the Google API will be removed, and a deep learning-based short-command recognition system will be used to increase the recognition accuracy. The Personalized Voice Activated Grasping System will also be tested with patients to improve the user interface.

CRediT authorship contribution statement

Yunfei Guo: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft. **Wenda Xu:** Validation, Formal analysis, Data curation, Writing – review & editing. **Sarthak Pradhan:** Validation, Formal analysis, Data curation, Writing – review & editing. **Cesar Bravo:** Conceptualization, Investigation, Writing – review & editing, Funding acquisition. **Pinhas Ben-Tzvi:** Conceptualization, Investigation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Yi Han, Yunhui Zhu, and Hailin Ren who provided valuable advice and greatly assisted in this research. The authors would also like to gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU.

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.

Appendix A. Multi-threading low-level control source code

https://github.com/yunfei96/Multi-threading-Exoskeleton-Glove-Control-Teensy-Arduino-.git

Appendix B. Speaker verification training source code

https://github.com/yunfei96/CVASV-Speaker-Verification-Training-.git

Appendix C. Cvasv hmi source code

https://github.com/yunfei96/Voice-Activation-and-Speaker-Verification-Command-System.git

References

[1] Refour EM, Sebastian B, Chauhan RJ, Ben-Tzvi P. A general purpose robotic hand exoskeleton with series elastic actuation. J. Mech. Robot. 2019;11(6):1–9. http://dx.doi.org/10.1115/1.4044543.

[2] Vanteddu T, Sebastian B, Ben-Tzvi P. Design optimization of RML glove for improved grasp performance. In: Proceedings of the ASME 2018 dynamic systems and control conference. vol. 1, 2018, p. 1–8. http://dx.doi.org/10.1115/DSCC2018-9004.

[3] Nilsson M, Ingvast J, Wikander J, Von Holst H. The soft extra muscle system for improving the grasping capability in neurological rehabilitation. In: 2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences, no. December. IEEE; 2012, p. 412–7. http://dx.doi.org/10.1109/IECBES.2012.6498090.

[4] Nilsen T, Hermann M, Eriksen CS, Dagfinrud H, Mowinckel P, Kjeken I. Grip force and pinch grip in an adult population: Reference values and factors associated with grip force. Scand. J. Occup. Ther. 2012;19(3):288–96. http://dx.doi.org/10.3109/11038128.2011.553687.

[5] Ma Z, Ben-Tzvi P. Design and optimization of a five-finger haptic glove mechanism. J. Mech. Robot. 2015;7(4):1–8. http://dx.doi.org/10.1115/1.4029437.

[6] Ma Z, Ben-Tzvi P, Danoff J. Hand rehabilitation learning system with an exoskeleton robotic glove. IEEE Trans Neural Syst Rehabil Eng 2016;24(12):1323–32. http://dx.doi.org/10.1109/TNSRE.2015.2501748.

[7] Ma Z, Ben-tzvi P, Member S. Sensing and force-feedback exoskeleton (SAFE) glove. In: Proceedings of the 2015 ASME IDETC/CIE, 39th mechanisms & robotics conference. Boston, MA; 2015.

[8] Lee BJB, Williams A, Ben-tzvi P, Member S. For rehabilitation and assistive applications. IEEE Trans Neural Syst Rehabil Eng 2018;26(8):1556–65.

[9] Lee J, Ben-Tzvi P. Design of a wearable 3-DOF forearm exoskeleton for rehabilitation and assistive purposes. In: Proceedings of the 2017 ASME international mechanical engineering congress and exposition conference. Tampa, Florida; 2017, http://dx.doi.org/10.1115/IMECE2017-71883.

[10] Chauhan RJ, Ben-Tzvi P. A series elastic actuator design and control in a linkage based hand exoskeleton. In: Proceedings of the ASME 2019 dynamic systems and control conference. Park City, Utah; 2019, http://dx.doi.org/10.1115/DSCC2019-8996.

[11] Chauhan R, Sebastian B, Member S, Ben-tzvi P, Member S. Exoskeleton glove control. IEEE Trans Hum-Mach Syst 2019;PP(1):1–10.

[12] Chauhan RJ, Ben-Tzvi P. Latent variable grasp prediction for exoskeletal glove control. In: Proceedings of the ASME 2018 dynamic systems and control conference. Atlanta, GA; 2018, http://dx.doi.org/10.1115/DSCC2018-8978.

[13] Vanteddu T, Ben-Tzvi P. Stable grasp control with a robotic exoskeleton glove. J. Mech. Robot. 2020;12(6):1–14. http://dx.doi.org/10.1115/1.4047724.

[14] Xu W, Pradhan S, Guo Y, Pinhas B-T, Bravo C. A novel design of a robotic glove system for paitients with brachial plexus injuries. In: Proceedings of the 2020 ASME IDETC/CIE, 44th mechanisms & robotics conference. St. Louis, MO; 2020.

[15] Feix T, Pawlik R, Schmiedmayer H-B, Romero J, Kragi D. A comprehensive grasp taxonomy. In: Proceedings of the robotics, science and systems conference: workshop on understanding the human hand for advancing robotic manipulation. 2009, p. 2–3.

[16] Li M, He B, Liang Z, Zhao CG, Chen J, Zhuo Y, et al. An attention-controlled hand exoskeleton for the rehabilitation of finger extension and flexion using a rigid-soft combined mechanism. Front. Neurorobotics 2019;13(May):1–13. http://dx.doi.org/10.3389/fnbot.2019.00034.

[17] Randazzo L, Iturrate I, Chavarriaga R, Leeb R, Millan JDR. Detecting intention to grasp during reaching movements from EEG. In: Proceedings of the annual international conference of the IEEE engineering in medicine and biology society. 2015-Novem, 2015, p. 1115–8. http://dx.doi.org/10.1109/EMBC.2015.7318561.

[18] Randazzo L, Iturrate In, Perdikis S, Millán JDR. Mano: A wearable hand exoskeleton for activities of daily living and neurorehabilitation. IEEE Robot Autom Lett 2018;3(1):500–7. http://dx.doi.org/10.1109/LRA.2017.2771329.

[19] Chowdhury A, Raza H, Meena YK, Dutta A, Prasad G. Online covariate shift detection-based adaptive brain-computer interface to trigger hand exoskeleton feedback for neuro-rehabilitation. IEEE Trans. Cogn. Dev. Syst. 2018;10(4):1070–80. http://dx.doi.org/10.1109/TCDS.2017.2787040.

[20] Baklouti M, Monacelli E, Guitteny V, Couvet S. Intelligent assistive exoskeleton with vision based interface. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5120 LNCS, 2008, p. 123–35. http://dx.doi.org/10.1007/978-3-540-69916-3_15.

[21] Kim D, Kang BB, Kim KB, Choi H, Ha J, Cho K-J, et al. Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view. Science Robotics 2019;4(26):eaav2949. http://dx.doi.org/10.1126/scirobotics.aav2949.

[22] Wang X, Tran P, Callahan SM, Wolf SL, Desai JP. Towards the development of a voice-controlled exoskeleton system for restoring hand function. In: Proceedings of the 2019 international symposium on medical robotics conference. Atlanta, GA: IEEE; 2019, p. 1–7. http://dx.doi.org/10.1109/ISMR.2019.8710195.

[23] Tripathy S, Panicker R, Shrey S, Naik R, Pachpore SS. Voice controlled upper body exoskeleton: A development for industrial application. 2020, arXiv.

[24] Densford F. Bionik labs integrates Amazon voice control tech into arke lower body exoskeleton. 2017.

[25] Lee JB. Development of intelligent exoskeleton grasping through sensor fusion and slip detection development of intelligent exoskeleton grasping [Ph.D. thesis], Virginia Polytechnic Institute and State University; 2018.

[26] Nagrani A, Chung JS, Xie W, Zisserman A. Voxceleb: Large-scale speaker verification in the wild. Comput Speech Lang 2020;60:101027. http://dx.doi.org/10.1016/j.csl.2019.101027, URL: https://www.sciencedirect.com/science/article/pii/S0885230819302712.

[27] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: EFficient convolutional neural networks for mobile vision applications. 2017, arXiv abs/170404861.

[28] Sueur J. A very short introduction to sound analysis for those who like elephant trumpet calls or other wildlife sound. 2014, p. 1–17.

[29] Balasingam MD, Kumar CS. Refining cosine distance features for robust speaker verification. In: Proceedings of the 2018 IEEE international conference on communication and signal processing, no. 1. 2018, p. 152–5. http://dx.doi.org/10.1109/ICCSP.2018.8524499.

[30] Nagraniy A, Chungy JS, Zisserman A. Voxceleb: A large-scale speaker identification dataset. In: Proceedings of the Annual Conference of the International Speech Communication Association, vol. 2017-Augus, 2017, p. 2616–20. http://dx.doi.org/10.21437/Interspeech.2017-950.

[31] Chung JS, Nagrani A, Zisserman A. Voxceleb2: Deep speaker recognition. In: Proceedings of the annual conference of the international speech communication association, vol. 2018-Septe, 2018, p. 1086–90. http://dx.doi.org/10.21437/Interspeech.2018-1929.

[32] Xie W, Nagrani A, Chung JS, Zisserman A. Utterance-level aggregation for speaker recognition in the wild. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing. IEEE; 2019, p. 5791–5.

[33] Liu W, Wen Y, Yu Z, Yang M. Large-margin softmax loss for convolutional neural networks. In: Proceedings of the 33rd international conference on international conference on machine learning, vol. 48, New York, NY, USA: JMLR.org; 2016, p. 507–16, URL: http://arxiv.org/abs/1612.02295.

[34] He S, Zhang A, Yan M. Voice and motion-based control system: Proof-of-concept implementation on robotics via Internet-of-Things technologies. In: Proceedings of the 2019 ACM southeast conference. 2019, p. 102–8. http://dx.doi.org/10.1145/3299815.3314440.

[35] El-emary IMM, Fezari M, Attoui H. Hidden Markov model/Gaussian mixture models (HMM/GMM) based voice command system: A way to improve the control of remotely operated robot arm TR45. Sci. Res. Essays 2011;6(2):341–50. http://dx.doi.org/10.5897/SRE10.759.

[36] Gómez JB, Ceballos A, Prieto F, Redarce T. Mouth gesture and voice command based robot command interface. In: Proceedings - IEEE international conference on robotics and automation, 2009, p. 333–8. http://dx.doi.org/10.1109/ROBOT.2009.5152858.

[37] Megalingam RK, Reddy RS, Jahnavi Y, Motheram M. ROS based control of robot using voice recognition. In: 2019 third international conference on inventive systems and control. 2019, p. 501–7. http://dx.doi.org/10.1109/ICISC44355.2019.9036443.

[38] Pleva M, Juhar J, Ondas S, Hudson CR, Bethel CL, Carruth DW. Novice user experiences with a voice-enabled human-robot interaction tool. In: 2019 29th International conference radioelektronika. 2019, p. 1–5. http://dx.doi.org/10.1109/RADIOELEK.2019.8733492.

[39] Guo S, Wang Z, Guo J, Fu Q, Li N. 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. 2018, p. 2300–5. http://dx.doi.org/10.1109/ICMA.2018.8484626.

Yunfei Guo received the B.S. degree in Electrical and Computer Engineering from Virginia Tech, USA in 2019. He is currently pursuing a Ph.D. degree at Virginia Tech, USA in the Robotics and Mechatronics Lab under the supervision of Prof. Pinhas Ben-Tzvi. His research interests include embedded systems and sensing, robotics control, and human–machine interfaces.

Wenda Xu received the B.S degree in Mechanical Engineering from Hunan University, Hunan, China, in 2016, and M.S. degree in Mechanical Engineering from Columbia University, New York, USA, in 2019. He is currently pursuing a Ph.D. degree at Virginia Tech, USA in the Robotics and Mechatronics Lab under the supervision of Prof. Pinhas Ben-Tzvi. His research interests include robotics design, artificial intelligence, and machine learning.

Sarthak Pradhan received his B.Tech. degree in Mechanical Engineering from Indian Institute of Technology, Bhubaneswar, India, in 2017. He is currently pursuing a master's degree in Mechanical Engineering at Virginia Tech, USA in the Robotics and Mechatronics Lab under the supervision of Prof. Pinhas Ben-Tzvi. His research interests include Exoskeletons, Humanoid Robots, Machine learning and Motion Planning and Localization of Robots.

Dr. Cesar Bravo completed an internship in general surgery and residency in orthopaedic surgery at University of Puerto Rico, following which he completed a fellowship in hand surgery at the Mayo Clinic, recognized as one of the world leaders in surgical brachial plexus injury repair. He specialized in hand and upper extremity surgery with a particular interest in problems of the elbow, peripheral nerve injuries, and hand and upper extremity trauma at Carilion Clinic since 2005. Dr. Bravo is certified by the American Board of Orthopaedic Surgery in orthopaedics and hand surgery. He is a Co-Director of Hand, Upper Extremity, and Microvascular Surgery within Carilion Clinic Orthopaedics.

Dr. Pinhas Ben-Tzvi (S'02–M'08–SM'12) received the B.S. degree (summa cum laude) in mechanical engineering from the Technion Israel Institute of Technology, Israel, and the M.S. and Ph.D. degrees in mechanical engineering from the University of Toronto, Canada. He is currently a Professor of Mechanical Engineering and Electrical and Computer Engineering, and the founding Director of the Robotics and Mechatronics Laboratory at Virginia Tech, Blacksburg, VA, USA. His current research interests include robotics and intelligent autonomous systems, human–robot interactions, robotic vision and visual servoing, machine learning, mechatronics design, systems dynamics and control, multi-robot and distributed systems, mechanism design and system integration, and novel sensing and actuation.

Dr. Ben-Tzvi is the recipient of the 2019 Virginia Tech Excellence in Teaching Award, 2018 Virginia Tech Faculty Fellow Award, the 2013 GWU SEAS Outstanding Young Researcher and Outstanding Young Teacher Awards, as well as several other honors and awards. Dr. Ben-Tzvi was a Technical Editor for the IEEE/ASME Transactions on Mechatronics, Associate Editor for ASME Journal of Mechanisms and Robotics, Associate Editor for IEEE Robotics and Automation Magazine, and an Associate Editor for the Int'l Journal of Control, Automation and Systems and served as an Associate Editor for IEEE ICRA 2013–2018. He is a Fellow of the American Society of Mechanical Engineers (ASME).