2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Voice-Driven Assistance and Resistance Modulation in a Soft Hip Exosuit Using a Transformer-based Speech Recognition Model

Enrica Tricomi[1*], Daniel Lindner[1], Xiaohui Zhang[1*], Luka Mišković[1] and Lorenzo Masia[1]

Abstract— **Intuitive human–robot interfaces are essential to increase usability and personalization in wearable robotic assistive technologies. However, most current systems rely on pre-programmed or sensor-driven strategies that offer limited active user control online. To address this limitation, we present a voice-driven control framework for a soft hip exosuit, enabling on-demand modulation of assistance and resistance via short spoken commands. The system combines a fully embedded transformer-based automatic speech recognition model (Whisper) with a gait-phase estimator to synchronize actuation with the user's motion. Users can switch between assistive and resistive modes and select discrete gain levels (low, medium, high). Experiments with six healthy participants demonstrate high recognition accuracy (95-100%) and low latency (∼9 ms). Metabolic measurements show that assistive commands reduced walking energy cost by 20.9±4.8% (LOW) and 9.7±5.5% (MEDIUM) relative to baseline, while resistive commands increased cost by 13.1±3.5% (MEDIUM) and 14.9±5.1% (HIGH). These results highlight the feasibility of intuitive, voice-driven modulation in wearable robotics.**

Index Terms— **Exosuits; Automatic Speech Recognition; Adaptive Assistance and Resistance; Wearable Robotics.**

I. Introduction

Wearable robotic assistive devices have the potential to enhance human locomotion by providing adaptive support tailored to individual needs [1], [2], [3]. Despite advances in actuation, sensing, and control, many systems remain pre-programmed or only semi-adaptive, often relying on predefined gait templates [4].

To improve adaptability, state-of-the-art controllers have leveraged kinematic and kinetic sensors, such as IMUs or force-sensing insoles, to estimate gait state and synchronize assistance with the user's motion [5], [6], [7]. Despite these advances, many systems remain restricted to a single operational mode, such as level walking [8], stair climbing [9], or resistance training [10], and generally lack intuitive mechanisms for reconfiguration. To extend beyond single-mode operation, computer vision or data driven estimation methods based on kinetics and kinematics have been integrated to detect changes in environments or user's state, enabling automatic switching between assistance modes [11], [12], [13], [14], [15]. While this improves context-awareness and responsiveness, mode selection is determined by the system rather than the user, leaving the wearer unable to directly adjust the type or intensity of support in real time.

To increase user agency, researchers have investigated direct-control interfaces on top of classic controllers such as manual buttons, sliders, or smartphone applications to select different support profiles [16]. These solutions provide a degree of intervention but often at the cost of manual effort or increased cognitive load. Enabling easier switching between modalities and intensity levels within a single platform therefore remains a key challenge, with the potential to improve personalization, reduce reliance on multiple devices, and expand the versatility of wearable assistive systems.

In this context, voice-driven control could represent a particularly intuitive and low-effort alternative for interacting with wearable assistive devices. By converting spoken commands into actionable control signals, voice interfaces could allow users to modulate assistance parameters, switch operational modes, or trigger specific functional behaviors without manual manipulation or interruption of the motor action [17]. Advances in automatic speech recognition (ASR) [18] and natural language processing [19] have made it possible to achieve robust transcription even in noisy environments, across different speakers and accents. These systems can also be implemented locally or on lightweight embedded platforms, ensuring low-latency and network-independent operation, which is critical for safety, responsiveness, and privacy in wearable robotics.

To date, voice control has been primarily investigated in humanoid robotics, where spoken commands are used to trigger walking patterns, postural adjustments, or task-specific behaviors [20]. In the context of wearable assistive robotics, instead, the application of voice-driven control remains largely unexplored. Such interfaces could provide a natural and continuous channel for user intent, where the user could directly influence device behavior in real time. This approach has the potential to enable context-aware modulation of support, such as switching between assistance modalities, adjusting the level of support, or activating task-specific functions, all while maintaining uninterrupted locomotion. Integrating the user as an active participant in the control loop could significantly enhance personalization, flexibility, and overall system usability, addressing a key limitation of current adaptive controllers and unlocking new possibilities for real-time, user-centered control of wearable robotic systems.

[1] E. Tricomi, D. Linder, X. Zhang, L. Mišković and L. Masia are with the Munich Institute for Robotics and Machine Intelligence (MIRMI), Department of Computer Engineering, School of Computation, Information and Technology, Technical University of Munich (TUM), Munich, Germany

This work was supported in part by the Istituto nazionale per l'assicurazione contro gli infortuni sul lavoro (INAIL) under grant agreement PR23-RR-P1 FeatherEXO, by the European Union through the SWAG Project under grant no. 101120408, and (partially) supported by the German Federal Ministry of Education and Research (BMBF) under the Robotics Institute Germany (RIG).

* corresponding authors: E. Tricomi enrica.tricomi@tum.de, X. Zhang xiaohui.zhang@tum.de

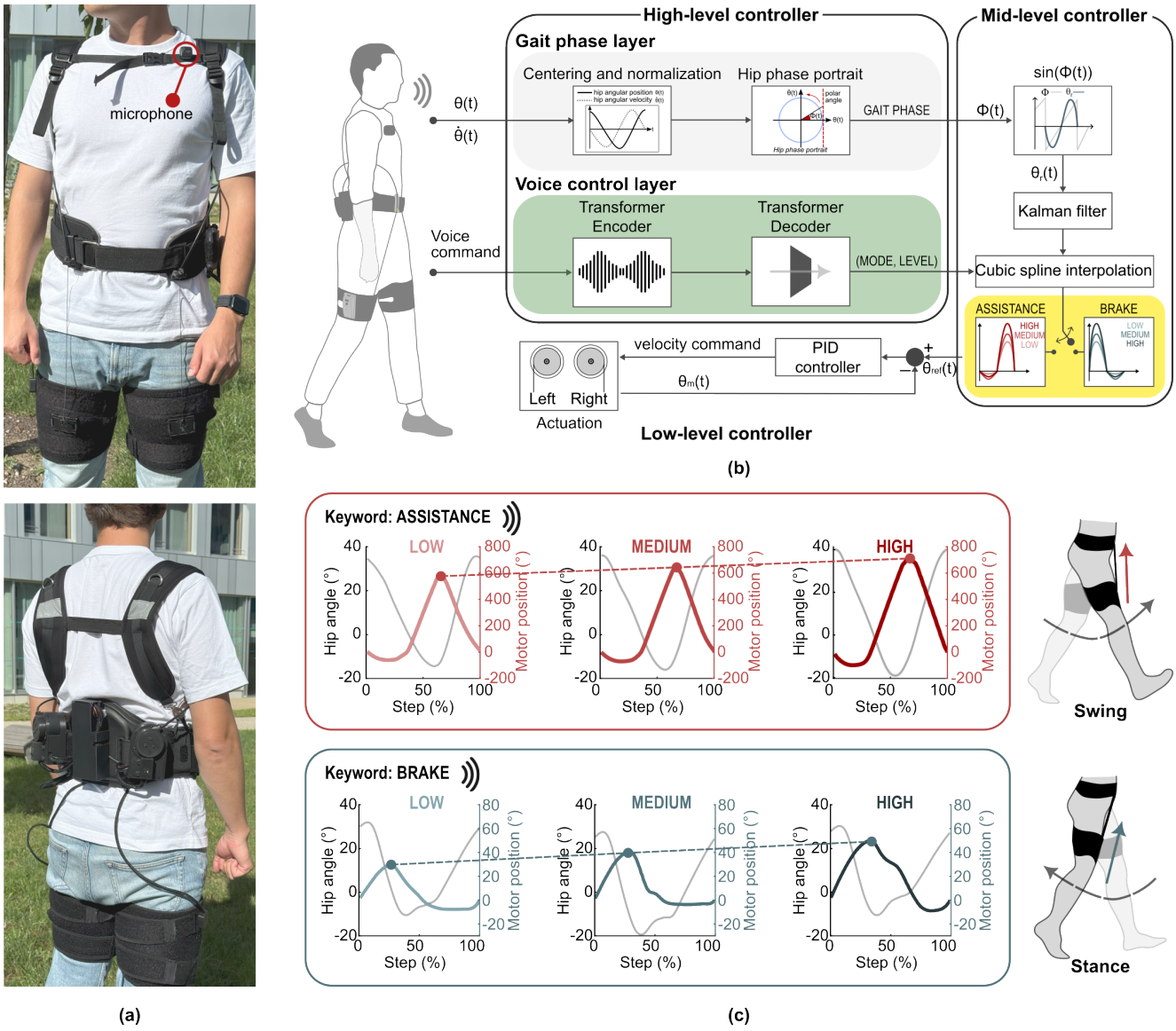

Fig. 1. *Hip exosuit design and voice-driven real-time control framework.* (a) Soft tendon-driven system for hip flexion assistance and resistance during walking. (b) Voice-driven real-time controller changing operational modality (assistance vs. resistance) and support level (low, medium, high) according to voice commands from the user. (c) User influence on motor commands modulation of assistance and resistance during walking according to the voice input.

To achieve the above mentioned goals, in this study, we introduce an embedded, voice-driven control framework for a soft hip exosuit, implemented using a fully local transformer-based automatic speech recognition model. The system enables online, user-driven modulation of the exosuit's operational modality, allowing users to switch between two modes, assistive and resistive, and adjust support across three discrete levels: low, medium, and high, corresponding to different assistance or resistance amplitudes. Experiments with six healthy participants evaluate system performance in terms of voice command recognition, latency, metabolic cost, and hip kinematics. The primary contribution of this work is to demonstrate the feasibility, responsiveness, and functional impact of embedding speech-driven commands within the controller of a wearable assistive device. The focus is on validating that voice input can be reliably decoded, integrated with gait-synchronized actuation, and used by the wearer to intentionally modulate assistance and resistance during walking.

II. Hip Exosuit

The voice-driven real-time controller was implemented on a fully-actuated tendon-driven soft hip exosuit programmed to provide assistance and resistance during walking.

The device (Fig.1-(a)) is composed of (i) two independent actuation modules, one per leg, serving as the core of the active support system, (ii) a pair of textile harnesses secured around the user's thighs, and (iii) a supportive waist belt which houses the actuation units and electronic components. Forces are transmitted to the wearer via external tendons (Black Braided Kevlar Fiber, KT5703-06, 2.2 kN maximum

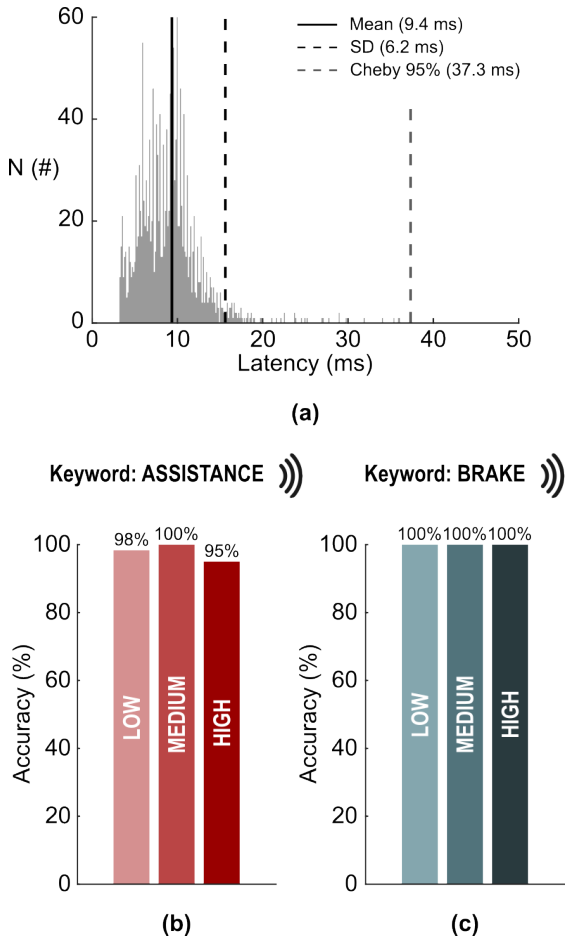

Fig. 2. *Latency and accuracy* (a) Mean latency computed as the time interval between the onset of a spoken command and the availability of the decoded output from the transformer-based model. (b-c) Recognition accuracy of the voice commands for mode and support level.

load, Loma Linda, CA, USA). The overall device mass is 2.4 kg, and it is battery powered (14.8 V, 3700 mA h, 45C).

Each actuation module incorporates a flat brushless motor (T-Motor, AK60-6, 24 V, 6:1 planetary gear-head reduction, Cube Mars actuator, TMOTOR, Nanchang, Jiangxi, China) that drives a ⌀35 mm pulley to wind the artificial tendon of the corresponding leg. Two Bowden cables (Shimano SLR, ⌀5 mm, Sakai, Ōsaka, Japan) connect the motors to proximal anchor points mounted on the waist belt, while distal tendon attachment points are 3D-printed and sewn onto the soft harnesses. This arrangement allows the motors to either assist hip flexion during swing or resist hip extension during stance, in accordance with the user's selected control mode.

Kinematic data are streamed to the controller from two inertial measurement units (IMU, Bosch, BNO055, Gerlingen, Germany), placed on the lateral sides of the thigh harnesses. Each IMU transmits hip angle and angular velocity information to the microcontrollers via Bluetooth Low Energy (BLE, Feather nRF52 Bluefruit, Adafruit).

Audio input for the voice-driven controller is recorded using a wireless microphone (DJI Mic Mini, Shenzhen, Guangdong, China), which the user clips near the head on clothing to optimize speech capture. This compact, lightweight device (10 g) operates on a 2.4 GHz wireless channel with a transmission range of up to 400 meters. The microphone incorporates dual omnidirectional capsules and digital signal processing to provide active noise cancellation, effectively reducing background noise and interference. Audio data are transmitted via Bluetooth to a receiver connected to the exosuit's embedded processor, ensuring low-latency, high-fidelity input for voice control.

The real-time control algorithm is distributed across multiple processors. A high-level controller handles speech transcription via the Whisper automatic speech recognition (ASR) model from OpenAI [21] and gait phase estimation from the hip IMU signals. Whisper is executed locally on an embedded processor (NVIDIA Jetson Nano, Santa Clara, CA, USA), ensuring rapid, network-independent audio processing. The gait phase estimator, the mid-level and low-level controllers run on Arduino MKR 1010 WiFi boards (Arduino, Ivrea, Italy) at 100 Hz, communicating with the high-level processor via serial.

Audio capture, transcription, and keyword parsing for voice commands are implemented in Python, while the remaining part of the control framework is implemented in MATLAB/Simulink (MathWorks, Natick, MA, USA).

III. Voice-driven real-time control framework

The real-time control framework of the hip exosuit is designed to change modality and modulate both *assistance* and *resistance* through short verbal commands, enabling the user to directly regulate the type and amplitude of support. The strategy combines automatic speech recognition (ASR) with a kinematics-based gait phase estimator, the latter building upon a previous work [22]. In this scheme, speech commands specify both the modality (assistance or resistance) and the gain level (low, medium, high), while the gait phase estimator ensures that the actuation of the artificial tendons is synchronized with the user's locomotion pattern.

The control architecture (Figure 1-(b)) is organized into three layers: (i) the high-level controller handles speech transcription and gait phase estimation; (ii) the mid-level controller uses this information to generate motor reference trajectories consistent with the commanded modality and gain; (iii) the low-level controller enforces trajectory tracking through a position-based feedback loop.

A. High-level controller

The high-level controller comprises two parallel processes: voice command recognition and gait phase estimation. In the following, these will be referred to as *voice control layer* and *gait phase layer*.

1) Voice control layer: The voice-based input interface is implemented using the Whisper automatic speech recognition (ASR) model from OpenAI [21], an open-source encoder–decoder transformer trained on 680,000 hours of multilingual and multitask supervised audio data. Unlike conventional ASR systems that depend on cloud-based processing, Whisper can run locally in lightweight variants (the smallest model `tiny.en` was used in this case) on CPUs

or GPUs, providing low-latency inference without requiring network access. This capability is particularly critical for embedded applications in wearable robotics, where a fully stand-alone implementation is preferred to ensure reliable operation independent of connectivity and ensuring privacy.

Architecturally, Whisper adopts a transformer encoder–decoder framework. The encoder operates on log-mel spectrograms of the input audio, extracting a contextual representation of the speech signal, while the decoder autoregressively generates text tokens conditioned on both the encoded audio features and the previously predicted tokens. In this respect, the decoder produces linguistic output sequentially, whereas the encoder provides grounding in acoustic information. This integration of large-scale sequence modeling with speech-specific encoding is what enables Whisper to generalize across speakers, accents, and noisy environments.

The implemented pipeline is organized into two cooperating threads. First, the *audio recorder* samples mono audio at 16 kHz in 4-second blocks and applies an amplitude gate to suppress near-silent segments. Second, the *transcriber* runs a faster-whisper model to convert the buffered audio into text transcriptions. These transcriptions are then parsed using deterministic regular-expression matchers of variable–level pairs (`ASSISTANCE|BRAKE + LOW|MEDIUM|HIGH`). The keyword `BRAKE` was intentionally adopted in place of `RESISTANCE`, since the latter was found to be frequently misinterpreted due to its phonetic similarity with `ASSISTANCE`. This substitution ensured more reliable keyword recognition while preserving a clear semantic distinction between assistive and resistive modes.

This architecture implements a keyword-based control scheme where each modality (assistance or resistance) is mapped to three discrete gain levels: low, medium, and high. The decoded command values are then forwarded to the mid-level controller, where they are integrated with the gait-phase estimation to generate the final motor command.

2) Gait phase layer: The gait phase of the user, $\phi(t)$, is estimated using hip kinematics measured by a single inertial sensor per leg. The hip angle $\theta(t)$ and angular velocity $\dot{\theta}(t)$ in the sagittal plane are first normalized and centered [23], which makes the hip phase portrait (angle vs. velocity) approximately circular and ensures a linear progression of the phase during each stride. The instantaneous gait phase is then computed as polar angle between angle and velocity:

$$\phi(t) = \arctan\left(\frac{\hat{\dot{\theta}}(t)}{\hat{\theta}(t)}\right)[\hat{\theta}(t) \neq 0] + \eta(t), \tag{1}$$

$$\eta(t) = sgn(\hat{\dot{\theta}}(t))\left(\pi[\hat{\theta}(t) < 0] + \frac{\pi}{2}[\hat{\theta}(t) = 0]\right), \tag{2}$$

where $\hat{\theta}(t)$ and $\hat{\dot{\theta}}(t)$ denote the normalized hip angle and velocity. The variable $\eta(t)$ corrects for the ambiguity introduced by the periodic back-and-forth motion. This provides a monotonically increasing phase variable $\phi(t)$ that reflects the progression of the gait cycle.

B. Mid-level controller

From the estimated gait phase, a preliminary motor trajectory is generated as a phase-dependent sinusoid:

$$\theta_r(t) = \sin(\phi(t)). \tag{3}$$

This raw signal, however, is sensitive to inertial sensor noise, particularly at heel strike or when the exosuit frame shifts. To improve robustness, $\theta_r(t)$ is filtered using a discrete Kalman filter [24], expressed as:

$$\hat{\mathbf{x}}_t = A\hat{\mathbf{x}}_{t-1} + K_t\big(\theta_r(t) - C\hat{\mathbf{x}}_{t-1}\big), \tag{4}$$

with the state vector $\hat{\mathbf{x}}_t = [\hat{\theta}_r(t), \hat{\dot{\theta}}_r(t)]^T$. The matrices

$$A = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \tag{5}$$

use $\Delta t = 0.01$ s. The process noise covariance Q and measurement noise covariance R are tuned as

$$Q = \begin{bmatrix} 0.02 & 0 \\ 0 & 0.02 \end{bmatrix}, \quad R = 0.75. \tag{6}$$

At this stage, the control modality uses cubic spline interpolation to determine how the output of this filter, $\hat{\theta}_r(t)$, is shaped into the final reference trajectory according to the modality. In *assistive mode*, the trajectory preserves the positive part of the sinusoidal waveform corresponding to hip flexion, while compressing the negative part, so that the motor pulls the tendon during swing to assist leg elevation. In *resistive mode*, the trajectory is inverted: the motor pulls the tendon during stance, impeding hip extension. This implementation exploits the tendon-driven transmission of the exosuit, in which the frontal cable arrangement allows the motors to pull only in one direction. As a result, assistive mode enhances swing initiation, whereas resistive mode opposes extension during stance.

Finally, the amplitude of the final trajectory depends on the gain level specified by the voice command. Gains are updated only at the rising zero-crossing of $\phi(t)$, which corresponds to the transition from stance to swing. This gating mechanism ensures that assistance or resistance is applied consistently within each gait cycle, avoiding abrupt intra-step changes.

Formally, the final motor reference trajectory is defined as

$$\theta_{\text{ref}}(t) = \begin{cases} S\Big[\, G_{\text{voice}} \cdot \max(\hat{\theta}_r(t), 0) + \\ G_{\text{voice}} \cdot \alpha \cdot \min(\hat{\theta}_r(t), 0)\,\Big], & \text{Assistive mode} \\ S\Big[\, G_{\text{voice}} \cdot \min(\hat{\theta}_r(t), 0) + \\ G_{\text{voice}} \cdot \alpha \cdot \max(\hat{\theta}_r(t), 0)\,\Big], & \text{Resistive mode} \end{cases} \tag{7}$$

where $\hat{\theta}_r(t)$ is the filtered sinusoidal gait-phase signal, G_{voice} is the gain level decoded from the user's voice command (`LOW`, `MEDIUM`, or `HIGH`), $\max(\cdot, 0)$ and $\min(\cdot, 0)$ select the positive and negative portions of the waveform, respectively, $\alpha \in [0, 1]$ is a compression factor applied to the suppressed portion, and $S[\cdot]$ denotes cubic spline interpolation. The gain G_{voice} is updated only at the rising

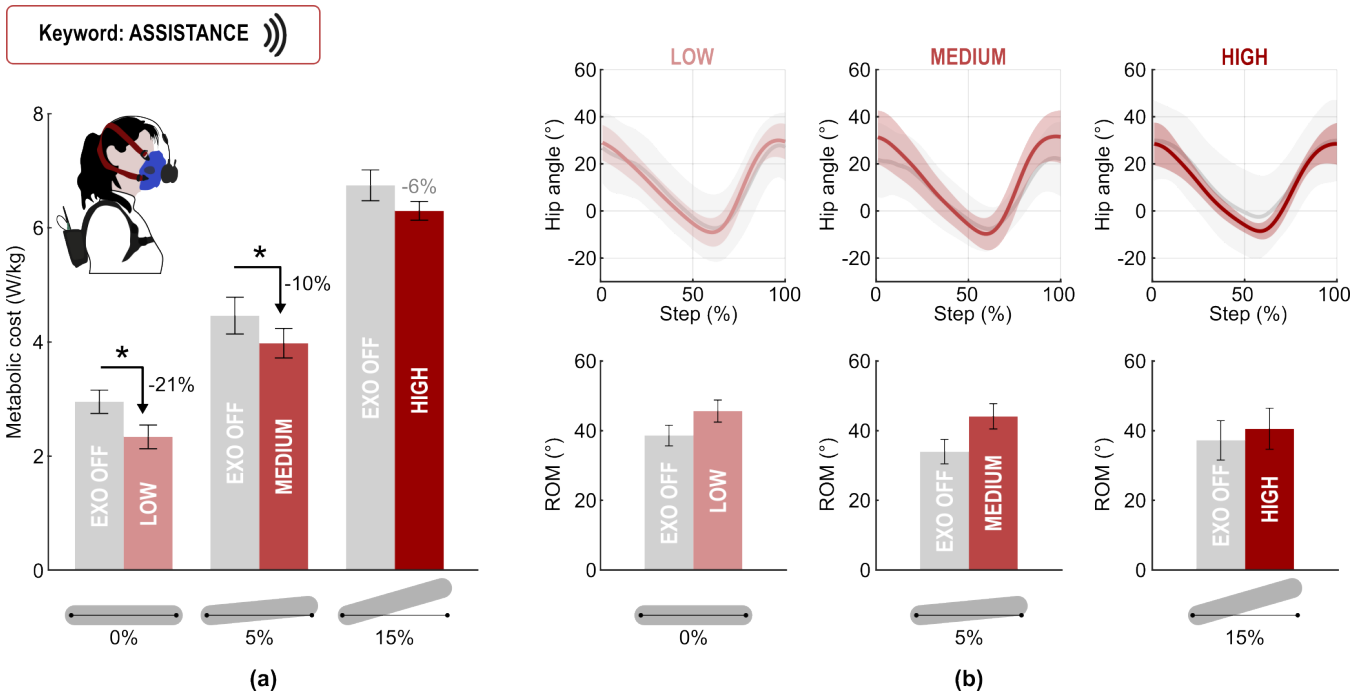

Fig. 3. *Metabolic and kinematic results for the assistance mode.* (a) Mean metabolic cost across subjects for the unassisted (exo off) and assisted conditions at different levels (low, medium, high). (b) Hip angle timeseries averaged across subjects (top row) and mean range of motion for the unassisted (exo off) and assisted conditions at different levels (low, medium, high). Tests were run at three treadmill inclinations (0%, 5%, 15%) corresponding to increasing levels of support as commanded by the user.

zero-crossing of the gait phase $\phi(t)$, which corresponds to the transition from stance to swing. This gating ensures that assistance or resistance is applied consistently within each gait cycle, preventing abrupt intra-step changes.

C. Low-level controller

The low-level controller closes the loop by comparing the actual motor position $\theta_m(t)$ with the commanded reference $\theta_{\text{ref}}(t)$. The position error drives a proportional–derivative (PD) controller with transfer function

$$Y(s) = \frac{K_p}{1 + K_d s}, \tag{8}$$

where K_p and K_d were tuned to ensure accurate and stable tracking of $\theta_{\text{ref}}(t)$ with low latency.

IV. Experiments

The performance of the voice-driven controller was evaluated in a study with six healthy adult volunteers (4 males and 2 females; age: 25.0 ± 2.2 years; weight: 60.9 ± 13.7 kg; height: 169.0 ± 9.9 cm; mean $\pm$ SD).

Participants represented different accents, including German, Italian, and Indian. All voice-control keywords were pronounced in English.

Written informed consent was obtained from each participant prior to the study. The experimental protocol adhered to the Declaration of Helsinki and was approved by the Ethics Committee of Heidelberg University (resolution S-313/2020).

A. Protocol design

Participants were instructed to walk on a treadmill at a constant speed of $4\,\text{km/h}$ while testing two different control modalities: *assistance* and *resistance*.

In the *assistance condition*, subjects completed three treadmill trials at different inclinations (0%, 5%, and 15%). Each trial lasted $4\,\text{min}$, resulting in a total duration of $12\,\text{min}$ per condition. At each change in inclination, participants were instructed to issue voice commands (`ASSISTANCE LOW`, `ASSISTANCE MEDIUM`, `ASSISTANCE HIGH`) so that the gain level increased with the increasing effort required for steeper inclines.

In the *resistance condition*, participants walked for $12\,\text{min}$ at constant 0% inclination. Every $4\,\text{min}$, they provided a new voice command (`BRAKE LOW`, `BRAKE MEDIUM`, `BRAKE HIGH`) to progressively increase the resistive loading and evaluate its effect on effort.

In the following, we will refer to the experimental conditions as `ASSISTANCE + LOW|MEDIUM|HIGH` and `BRAKE + LOW|MEDIUM|HIGH`, corresponding to real-time voice-driven modulation in assistive and resistive modes, respectively. The protocol was additionally repeated with the exosuit deactivated (`EXO OFF`) for baseline comparison.

The order of conditions was randomized across participants to minimize learning and fatigue effects. Rest intervals of at least $20\,\text{min}$ were provided between trials.

B. Data acquisition

Task effort was quantified through the metabolic cost of walking, measured with a portable gas analyzer (K5,

COSMED, Rome, Italy) that continuously monitored oxygen uptake and carbon dioxide production. At the beginning of each session, participants completed a 4 min quiet standing trial with spontaneous breathing to establish an individual baseline. This baseline metabolic rate was then subtracted from the walking measurements to obtain the net locomotion cost for each condition.

In addition to metabolic measurements, timestamps of decoded voice commands from the *voice control layer*, hip kinematics from the IMU sensors embedded in the thigh harnesses, and motor actuation data from the exosuit were recorded.

C. Data analysis

1) Voice command accuracy and latency: The performance of the *voice control layer* was assessed in terms of end-to-end latency and recognition accuracy. Latency was defined as the time interval between the onset of a spoken command and the availability of the decoded output from the transformer-based model. Operationally, when the microphone captured a 4-second audio segment exceeding a predefined volume threshold, the system registered a timestamp at the end of the segment. A second timestamp was recorded upon completion of the speech-to-text decoding process. Latency was computed as the difference between these two timestamps. All processed audio segments from the all treadmill trials (both assistance and resistance) were included in the analysis, regardless of whether they contained noise or background speech.

Accuracy was instead calculated for each verbal command as the ratio of correctly recognized instances of a given keyword pair (`ASSISTANCE|BRAKE + LOW|MEDIUM|HIGH`) to the total number of predicted commands under the tested condition.

2) Metabolic cost: Metabolic energy expenditure was estimated from oxygen uptake and carbon dioxide production. Net metabolic cost was calculated according to the equations of Péronnet and Massicotte [25], normalized to body weight, and corrected by subtracting the baseline value obtained during quiet standing. In the assistive trial, paired comparisons (exo off vs. assisted condition) were conducted for each treadmill inclination (0%, 5%, and 15%) to examine whether the level of assistance could help mitigate the increase in energetic demand associated with greater task difficulty. In the resistive trial, data were collected and analyzed only at 0% inclination, as the objective was to demonstrate that progressively higher resistance levels were associated with increasing energetic cost.

3) Hip kinematics: Hip joint kinematics were segmented into individual gait cycles. The primary outcome was the hip range of motion (ROM), which was compared across conditions.

D. Statistical analysis

Data normality was assessed using the Shapiro–Wilk test ($\alpha = 0.05$). Since all distributions met normality assumptions, parametric testing was applied. A linear mixed-effects model was implemented (MATLAB, MathWorks Inc., Natick, MA, USA) to analyze both metabolic and kinematic outcomes. The model included *condition* (`EXO OFF`, `ASSISTANCE + LOW|MEDIUM|HIGH`, `BRAKE + LOW|MEDIUM|HIGH`) as a fixed-effect categorical variable, while *participant* was modeled as a random-effect factor to account for inter-subject variability. Statistical significance was set at $p < 0.05$ for all tests.

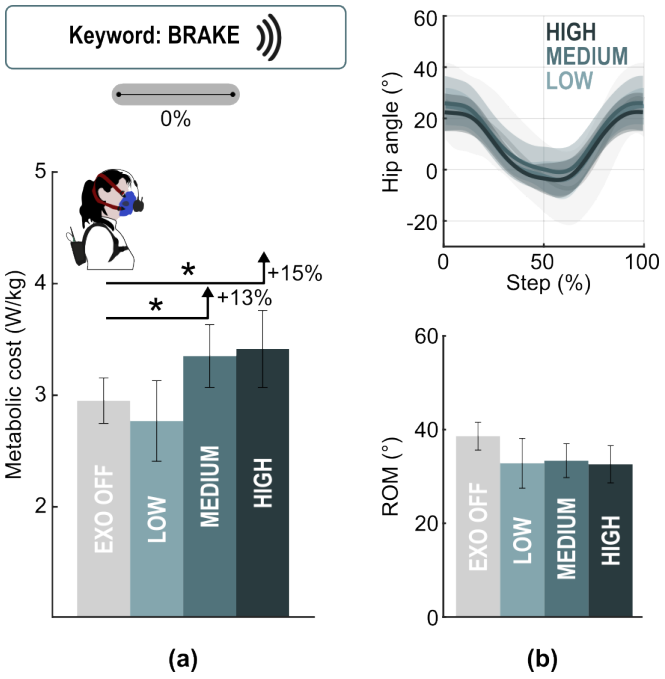

Fig. 4. *Metabolic and kinematic results for the resistance mode.* (a) Mean metabolic cost across subjects for the unassisted (exo off) and resistive conditions at different levels (low, medium, high). (b) Hip angle timeseries averaged across subjects (top row) and mean range of motion for the unassisted (exo off) and resistive conditions at different levels (low, medium, high). Tests were run at constant treadmill inclinations (0%).

V. Results

A. Voice command accuracy and latency

Voice command latency and accuracy for operating the exosuit controller are reported in Fig.2. The mean latency between voice input and transcription by the ASR model (Fig.2-(a)) was 9.4 ± 6.2 ms (mean$\pm$SD) across all subjects and trials.

In the assistive trial (Fig.2-(b)), recognition accuracy averaged 98% for the `ASSISTIVE LOW` command, 100% for the `ASSISTIVE MEDIUM` command, and 95% for the `ASSISTIVE HIGH` command.

In the resistive trial (Fig. 2-(c)), recognition accuracy was 100% for all `BRAKE` commands, irrespective of resistance level (`LOW`, `MEDIUM`, or `HIGH`).

B. Metabolic cost

For the assistive trial (Fig. 3-(a)), the metabolic cost of walking across the three treadmill inclinations (0%, 5%, and 15%) was generally lower in the assisted conditions compared to the `EXO OFF` condition. Specifically, the

ASSISTANCE LOW command reduced metabolic cost by $-20.9 \pm 4.8\%$ relative to EXO OFF ($p = 0.002$), while the ASSISTANCE MEDIUM command yielded a reduction of $-9.7 \pm 5.5\%$ ($p = 0.04$). In contrast, the ASSISTANCE HIGH condition did not show a statistically significant difference compared to EXO OFF ($p = 0.1$), although a trend toward reduced metabolic cost was observed ($-5.9 \pm 4.1\%$).

For the resistive trial (Fig. 4-(a)) at 0% treadmill inclination, metabolic cost increased significantly under the BRAKE MEDIUM and BRAKE HIGH conditions relative to EXO OFF, by $+13.1 \pm 3.5\%$ ($p = 0.007$) and $+14.9 \pm 5.1\%$ ($p = 0.02$), respectively. The BRAKE LOW condition, however, did not differ significantly from *Exo off* ($p = 0.7$).

C. Hip kinematics

Hip range of motion (ROM) did not reach statistical significance in either the assistive trial (Fig. 3-(b)) or the resistive trial (Fig. 4-(b)); however, a clear trend is evident across participants. In each panel of Figs. 3-(b) and 4-(b), the top plots illustrate subject-averaged timeseries for the three conditions (no assistance, assistance, resistance), while the bottom plots show the corresponding mean ROM values. Data represent the average between the right and left legs, as no side-related differences were found.

During the assistive trial, hip ROM exhibited a consistent increasing trend across all treadmill inclinations, with an average rise of approximately $+19^\circ$ compared to the EXO OFF condition, although this effect did not reach statistical significance ($p > 0.05$). Conversely, in the resistive trial, hip ROM showed a decreasing trend under the resistive mode, with an average reduction of about -16° relative to EXO OFF ($p > 0.05$).

VI. Discussion and conclusion

Lower-limb wearable robotic devices have long aimed to enhance human locomotion through adaptive and personalized assistance. While recent advances have largely focused on improving control algorithms and enhancing environmental responsiveness, most systems still provide limited shared control, rarely integrating the user's intent directly into the loop. State-of-the-art adaptive controllers typically rely on sensor-based gait phase estimation [11], [6], [1]. These approaches improve timing and responsiveness relative to the user's motion but generally do not allow on-demand modulation of operational mode or support level, limiting the degree of personalization achievable in real time.

The primary contribution of this work is the demonstration of an embedded, voice-driven control framework that enables real-time modulation of both assistance and resistance in a soft hip exosuit. To the best of our knowledge, this is the first instance of a system that integrates switching between operational modalities (assistance vs. resistance) and discrete support levels (low, medium, high) within a single platform. By enabling users to adjust both the type and intensity of interaction through simple spoken commands, the framework advances the development of user-centered wearable robotics and shared control strategies.

A distinguishing feature of our system is its fully offline operation on embedded hardware, eliminating the need for external servers or continuous network connectivity. This ensures reliable and private processing of voice commands and enhances system robustness in real-world applications. However, the offline approach also imposes constraints on model complexity and computational resources, which may limit transcription accuracy in acoustically challenging or noisy environments compared to cloud-based alternatives. Future work should explore lightweight hybrid strategies that maintain offline responsiveness while leveraging occasional cloud-assisted refinement for more complex commands.

The experimental protocol was designed to evaluate the system under both assistive and resistive conditions while enabling real-time voice-based modulation of support levels. Assistive trials were conducted at varying treadmill inclinations to simulate changing locomotor demands and test whether participants could dynamically adjust support in response to increased mechanical work. Resistive trials were performed at a single inclination to assess the system's ability to progressively scale resistive loading. In this initial evaluation, changes in assistance level were intentionally coupled to treadmill inclination to create controlled and repeatable variations in locomotor demand. While this design limits user-driven exploration of preferred assistance levels, it allows a structured assessment of whether voice commands can reliably trigger meaningful changes in device behavior under increasing task difficulty. Allowing participants to freely explore assistance levels, perform rapid or repeated mode switches, or select preferred gains at fixed task conditions may better capture the practical advantages of voice-based interaction. Such paradigms were outside the scope of the present feasibility study but represent an important direction for future work.

Results demonstrate high voice command recognition accuracy (95-100%) (Fig. 2) with low latency ($\sim$9 ms), supporting responsive and reliable user control. Notably, recognition accuracy in the resistive trial reached 100% across all levels, which may be partly explained by the choice of the keyword BRAKE: a short, common monosyllabic word that is easier for the ASR model to recognize compared to the longer and less frequent term ASSISTANCE.

Metabolic measurements indicate that assistive commands significantly reduced the energetic cost of walking at low and medium levels. A trend toward reduction is evident also at the highest assistance level, though this did not reach statistical significance, likely due to the limited sample size (Fig. 3-(a)). Resistive commands produced the expected increase in metabolic demand, except at the lowest resistance level, where effects were less clear, potentially reflecting both the modest loading and the mechanical constraints of implementing resistance in a tendon-driven exosuit (Fig. 4-(a)). Hip range of motion did not reach statistical significance but showed consistent trends aligned with the commanded modality, suggesting that voice-driven modulation can influence gait mechanics in accordance with user input. Collectively, these findings support the feasibility and functional

relevance of voice-driven control for modulating both type and intensity of wearable assistance in real time.

This study has several limitations. First, it does not include a direct comparison between voice-driven control and alternative interaction modalities such as buttons, sliders, or sensor-based automatic switching. Consequently, the present work does not evaluate whether speech input reduces cognitive load, enhances perceived intuitiveness, or offers advantages in reliability relative to other interfaces. Although recognition accuracy exceeded 95%, manual controls would likely achieve near-perfect input fidelity, and the impact of occasional recognition failures on user effort remains an open question. Second, the evaluation was preliminary and involved only six healthy young adults, which limits statistical power and generalizability. The results should therefore be interpreted as a proof of concept rather than evidence of broad effectiveness. Future studies should include larger and more diverse populations, particularly older adults or individuals with motor impairments, for whom hands-free interaction may be especially beneficial. Moreover, experiments were conducted under controlled treadmill conditions, and user experience metrics such as perceived control, comfort, and cognitive load were not formally assessed, leaving the subjective impact of voice-driven interaction unquantified. Future work should extend the evaluation to overground and outdoor walking, incorporate subjective usability measures, and explicitly compare voice-based control with alternative interfaces using both objective performance metrics and user feedback. Finally, the implemented command set was limited to two modalities and three discrete intensity levels. While sufficient for this feasibility study, the proposed framework can be readily extended to a richer command vocabulary, enabling more nuanced and personalized control strategies.

The integration of voice-driven control into wearable exosuits highlights the potential for intuitive, flexible, and user-centered robotic assistance. Providing users with direct control over both support modality and intensity represents a meaningful advancement toward more adaptive, personalized, and widely adoptable wearable robotic systems.

REFERENCES

[1] M. Xiloyannis, R. Alicea, A.-M. Georgarakis, F. L. Haufe, P. Wolf, L. Masia, and R. Riener, "Soft robotic suits: State of the art, core technologies, and open challenges," *IEEE Transactions on Robotics*, 2021.

[2] J. M. Cha, J. Hong, J. Yoo, and D.-W. Rha, "Wearable robots for rehabilitation and assistance of gait: A narrative review," *Annals of rehabilitation medicine*.

[3] A. J. Young and D. P. Ferris, "State of the art and future directions for lower limb robotic exoskeletons," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 2, pp. 171–182, 2016.

[4] M. R. Tucker, J. Olivier, A. Pagel, H. Bleuler, M. Bouri, O. Lambercy, J. d. R. Millán, R. Riener, H. Vallery, and R. Gassert, "Control strategies for active lower extremity prosthetics and orthotics: a review," *Journal of neuroengineering and rehabilitation*, vol. 12, no. 1, pp. 1–30, 2015.

[5] X. Zhang, E. Tricomi, F. Missiroli, N. Lotti, and L. Masia, "Real-time assistive control via imu locomotion mode detection in a soft exosuit: An effective approach to enhance walking metabolic efficiency," *IEEE/ASME Transactions on Mechatronics*, vol. 29, no. 3, pp. 1797–1808, 2023.

[6] D. D. Molinaro, I. Kang, and A. J. Young, "Estimating human joint moments unifies exoskeleton control, reducing user effort," *Science robotics*, vol. 9, no. 88, p. eadi8852, 2024.

[7] Y. Qian, Y. Wang, C. Chen, J. Xiong, Y. Leng, H. Yu, and C. Fu, "Predictive locomotion mode recognition and accurate gait phase estimation for hip exoskeleton on various terrains," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6439–6446, 2022.

[8] L. H. Sloot, L. M. Baker, J. Bae, F. Porciuncula, B. F. Clément, C. Siviy, R. W. Nuckols, T. Baker, R. Sloutsky, D. K. Choe *et al.*, "Effects of a soft robotic exosuit on the quality and speed of overground walking depends on walking ability after stroke," *Journal of neuroengineering and rehabilitation*, vol. 20, no. 1, p. 113, 2023.

[9] H. D. Lee, H. Park, B. Seongho, and T. H. Kang, "Development of a soft exosuit system for walking assistance during stair ascent and descent," *International Journal of Control, Automation and Systems*, vol. 18, no. 10, pp. 2678–2686, 2020.

[10] K. Swaminathan, S. Park, F. Raza, F. Porciuncula, S. Lee, R. W. Nuckols, L. N. Awad, and C. J. Walsh, "Ankle resistance with a unilateral soft exosuit increases plantarflexor effort during pushoff in unimpaired individuals," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, no. 1, p. 182, 2021.

[11] E. Tricomi, N. Lotti, F. Missiroli, X. Zhang, M. Xiloyannis, T. Muller, S. Crea, E. Papp, J. Krzywinski, N. Vitiello *et al.*, "Underactuated soft hip exosuit based on adaptive oscillators to assist human locomotion," *IEEE Robotics and Automation Letters*, 2021.

[12] B. Laschowski, W. McNally, A. Wong, and J. McPhee, "Environment classification for robotic leg prostheses and exoskeletons using deep convolutional neural networks," *Frontiers in Neurorobotics*, vol. 15, p. 730965, 2022.

[13] A. G. Kurbis, B. Laschowski, and A. Mihailidis, "Stair recognition for robotic exoskeleton control using computer vision and deep learning," in *2022 International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2022, pp. 1–6.

[14] D. D. Molinaro, K. L. Scherpereel, E. B. Schonhaut, G. Evangelopoulos, M. K. Shepherd, and A. J. Young, "Task-agnostic exoskeleton control via biological joint moment estimation," *Nature*, vol. 635, no. 8038, pp. 337–344, 2024.

[15] J. Lin, G. C. Thomas, N. V. Divekar, V. Peddinti, and R. D. Gregg, "A modular framework for task-agnostic, energy shaping control of lower limb exoskeletons," *IEEE Transactions on Control Systems Technology*, 2024.

[16] J. R. Diego, D. W. C. Martinez, G. S. Robles, and J. R. C. Dizon, "Development of smartphone-controlled hand and arm exoskeleton for persons with disability," *Open Engineering*, vol. 11, no. 1, pp. 161–170, 2020.

[17] 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.

[18] S. J. Arora and R. P. Singh, "Automatic speech recognition: a review," *International Journal of Computer Applications*, vol. 60, no. 9, 2012.

[19] Q. Xu, Z. Feng, C. Gong, X. Wu, H. Zhao, Z. Ye, Z. Li, and C. Wei, "Applications of explainable ai in natural language processing," *Global Academic Frontiers*, vol. 2, no. 3, pp. 51–64, 2024.

[20] T. Asfour, M. Waechter, L. Kaul, S. Rader, P. Weiner, S. Ottenhaus, R. Grimm, Y. Zhou, M. Grotz, and F. Paus, "Armar-6: A high-performance humanoid for human-robot collaboration in real-world scenarios," *IEEE Robotics & Automation Magazine*, vol. 26, no. 4, pp. 108–121, 2019.

[21] J. R. Batista, *Learn OpenAI Whisper: Transform your understanding of GenAI through robust and accurate speech processing solutions*. Packt Publishing Ltd, 2024.

[22] E. Tricomi, F. Missiroli, M. Xiloyannis, N. Lotti, X. Zhang, M. Stefanakis, M. Theisen, J. Bauer, C. Becker, and L. Masia, "Soft robotic shorts improve outdoor walking efficiency in older adults," *Nature Machine Intelligence*, vol. 6, no. 10, pp. 1145–1155, 2024.

[23] D. Quintero, D. J. Lambert, D. J. Villarreal, and R. D. Gregg, "Real-time continuous gait phase and speed estimation from a single sensor," in *2017 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2017, pp. 847–852.

[24] G. Welch, G. Bishop *et al.*, "An introduction to the kalman filter," 1995.

[25] F. Peronnet, D. Massicotte *et al.*, "Table of nonprotein respiratory quotient: an update," *Can J Sport Sci*, vol. 16, no. 1, pp. 23–29, 1991.