electronics

MDPI

Article

A Voice User Interface on the Edge for People with Speech Impairments

Davide Mulfari * and Massimo Villari

MIFT Department, University of Messina, 98122 Messina, Italy; mvillari@unime.it
* Correspondence: dmulfari@unime.it

Abstract: Nowadays, fine-tuning has emerged as a powerful technique in machine learning, enabling models to adapt to a specific domain by leveraging pre-trained knowledge. One such application domain is automatic speech recognition (ASR), where fine-tuning plays a crucial role in addressing data scarcity, especially for languages with limited resources. In this study, we applied fine-tuning in the context of atypical speech recognition, focusing on Italian speakers with speech impairments, e.g., dysarthria. Our objective was to build a speaker-dependent voice user interface (VUI) tailored to their unique needs. To achieve this, we harnessed a pre-trained OpenAI's Whisper model, which has been exposed to vast amounts of general speech data. However, to adapt it specifically for disordered speech, we fine-tuned it using our private corpus including 65 K voice recordings contributed by 208 speech-impaired individuals globally. We exploited three variants of the Whisper model (small, base, tiny), and by evaluating their relative performance, we aimed to identify the most accurate configuration for handling disordered speech patterns. Furthermore, our study dealt with the local deployment of the trained models on edge computing nodes, with the aim to realize custom VUIs for persons with impaired speech.

Keywords: automatic speech recognition; whisper; dysarthria; atypical speech; transformer; edge; assistive technology; AI

Citation: Mulfari, D.; Villari, M. A Voice User Interface on the Edge for People with Speech Impairments. *Electronics* **2024**, *13*, 1389. https://doi.org/10.3390/electronics13071389

Academic Editor: Martin Reisslein

Received: 11 March 2024
Revised: 2 April 2024
Accepted: 5 April 2024
Published: 7 April 2024

1. Introduction

Human speech serves as a fundamental channel for interpersonal communication and expression of ideas. As technology advances, voice user interfaces (VUIs) have gained prominence, offering an alternative to traditional input methods like keyboards and touchscreens. These VUIs leverage automatic speech recognition (ASR) services to process spoken commands, enabling natural interactions with machines across diverse application scenarios, including smart home automation. However, despite their widespread adoption, current speech recognition systems face a critical challenge: inclusivity. Millions of users with speech impairments, such as dysarthria, encounter barriers when attempting to benefit from voice-controlled devices. Dysarthria, a complex neuromotor voice disorder can be either acquired or congenital. It manifests through abnormalities in the strength, speed, range, steadiness, tone, or accuracy of movements required for speech production [1]. The negative impact on speech intelligibility is profound, particularly when coupled with severe movement disorders resulting from conditions like cerebral palsy or progressive diseases. As a result, the speech is atypical and exhibits distinct features, including imprecise pitch pauses during consonant and vowel production, leading to a lack of clarity in phonemes. Unfortunately, conventional ASR systems rely on auditory cues that may be obscured by these deviations. Furthermore, the effects of speech disability vary significantly among individuals, making the speech variations observed in dysarthric persons considerably greater than those found in typical speech. In light of these complexities, it is evident that existing ASR models demonstrate poor performance when confronted with impaired speech. Several authors have highlighted that virtual home assistants do not perform at a

sufficient level for people with a speech impairment, in fact the more severe the case, the worse the performance. Detailed investigations can be found in [2–4].

One of the major concerns in the development of speech recognition solutions for pathological speech is the lack of available data [5], which accelerates the insufficient generalization of ASR performance among speakers. Consequently, disordered speech recognition ASR development can be rephrased as the challenge of building a model with a limited amount of data nowadays. To this end, in prior studies, a transfer learning approach was explored where an acoustic model was first trained with a large healthy speech corpus and then fine-tuned with a dataset including pathological speech [6]. Although this approach showed positive outcomes, it is still far away from being competitive with ASR for speech without disorders.

The present study attempted to fill this gap by proposing, in the domain of atypical ASR, a fine-tuning approach that exploits a state-of-the-art (SOTA) speech recognition architecture, namely, OpenAI's Whisper. It employs an encoder–decoder Transformer structure leading to the creation of a supervised learning-based ASR system, which uses large amounts of labeled audio data, specifically 680K hours. The model uses weakly supervised pre-training beyond English-only speech recognition to be multilingual and for multitasking, showing great performance on different multilingual speech datasets [7]. As a sequence-to-sequence model, Whisper maps a sequence of audio spectrogram features to a chain of characters, while it applies a speech vision approach helping to address the limitations of variations in phonemes, their labeling, and the scarcity of impaired audio data, as motivated by recent scientific contributions [8,9]. In this study, we propose to fine-tune Whisper on our private corpus of Italian atypical speech encompassing 65 K single speech recordings. These data have been authored by 208 anonymized individuals with various conditions causing a speech disorder. We adopt a pure speaker dependent methodology wherein three different Whisper variants—small, base, tiny—are trained to spot precise and unique keywords belonging to our closed ASR dictionary, including 79 distinct elements, i.e., Italian words. Owing to the collaboration of 16 participants with diverse grades of speech impediments, the effectiveness of our approach was evaluated in terms of word recognition accuracy (WRA) to investigate the performance of Whisper across various speech disorders. Additionally, we explored the feasibility of running inference tasks on edge computing nodes, specifically single board computers. This investigation contributes insights toward the development of a local Voice User Interface (VUI) tailored for disordered speech, which is an area of high demand in assistive technology [10].

To the best of our knowledge, the utilization of fine-tuning techniques for automatically recognizing disordered Italian speech has not been thoroughly explored to date. Consequently, the main contributions of the present study are summarized below:

- Fine-tuning three distinct variants of OpenAI's Whisper using our proprietary collection of Italian atypical speech;
- Evaluating their relative accuracy in keyword spotting, with the valuable input of selected individuals who have speech disabilities;
- Harnessing these fine-tuned models to create a voice user interface tailored for Italian individuals who experience dysarthria and other speech disorders.

The rest of this article is structured as follows. Section 2 presents related works. Details about our methodology are provided in Section 3. Section 4 discusses our experimental outcomes, and Section 5 summarizes the conclusion of the study.

2. Related Works

The numerous challenges of automated speech recognition in the presence of atypical patterns have gained the attention of both industry and research communities [11–16]. To overcome these issues, the collaboration between diverse researchers and actors plays a crucial role. In this regard, the "Speech Accessibility Project" is a collaborative initiative led by the University of Illinois Urbana-Champaign (UIUC) in partnership with major technology companies, including Amazon, Apple, Google, Meta, and Microsoft. Its main

objectives is to enhance current voice recognition technology to be more inclusive and useful for individuals with diverse speech patterns and disabilities. In this context, the adoption of fine-tuning-based approaches is of paramount importance. Generally, within the framework of deep learning, fine-tuning refers to the process of adjusting a pre-trained model on a specific task or dataset. In ASR, such a technique helps adapt the model to specific accents, languages, or speakers' characteristics, while the model's parameters are adjusted using a smaller, task-specific dataset (fine-tuning data).

Nowadays, fine-tuning on accents or atypical populations significantly improves performance and, at the same time, allows ASR systems to handle variations in speech patterns more effectively [17]. Positive results in the Whisper model's fine-tuning have been documented in the domain of child speech recognition [18], as well as in dysarthric speech recognition, where Rathod et al. highlighted a WRA of 59% using a block of 155 keywords belonging to the English UA-Speech corpus [19]. Furthermore, a Whisper application in a rehabilitation scenario involving patients with post-stroke aphasia is proposed in [20].

Different studies have focused on the possible utilization of self-supervised learning [21] (SSL) approaches in the presence of atypical speech to address the main challenge of speech data scarcity [22–25]. For example, Wang et al. [26] explored the advantages of pre-trained mono and cross-lingual speech representations for the spoken language understanding of Dutch dysarthric speech, and Hu et al. [27] investigated a series of approaches to integrate domain-adapted SSL pre-trained models into time delay neural networks and conformer ASR systems for elderly and impaired speech recognition by working on the UA-Speech and the Dementia Pitt corpora.

As shown in Table 1, the application of fine-tuning methods in the automated recognition of Italian disordered speech remains unexplored. To fill this gap, this study performed the following:

- Fine-tuned three different OpenAI's Whisper variants on our private corpus of Italian atypical speech;
- Measured the relative performance in terms of accuracy in keyword spotting, owing to the collaborative effort of selected individuals with a speech disability;
- Leveraged the fine-tuned models toward the development of a voice user interface for users with atypical voices who speak Italian.

Table 1. Recent studies on the utilization of fine-tuning approaches in impaired speech recognition.

Reference	Language	Method
[19]	English (UaSpeech corpus)	Whisper and Bi-LSTM classifier model
[20]	English (SONIVA corpus)	Whisper large model
[23]	English (UaSpeech) German (private corpus)	Fne-tuning Wav2Vec2 using fMLLR features
[24]	English (UaSpeech) Japanese (ELSpeech corpus)	Wav2Vec2 + Wav2LM
[27]	English (UaSpeech and DementiaBank corpus)	SSL pre-trained Wav2Vec2 with hybrid TDNN and Conformer
[26]	Dutch (private database)	Various Wav2Vec2 and Whisper variants
[28]	English and Spanish (AphasiaBank database)	Wav2Vec2 XLSR-53 model

3. Methodology

This section presents our speaker-dependent methodology designed for the automated recognition of unique and precise words spoken by individuals with speech disorders, e.g., dysarthria, who are Italian speakers. In the context of atypical speech recognition,

our proposed approach focuses on keyword spotting tasks rather than spoken language understanding. Indeed, for many individuals living with speech disorders, single words serve as a convenient natural voice production method, helping to mitigate difficulties in breath control and speech coordination. Consequently, isolated word recognizers play a crucial role in minimizing dysarthric ASR errors [29].

Given the lack of adequate corpora containing Italian disordered speech samples, one of our primary objectives was to create such a database. We leveraged findings from previous scientific contributions and successfully utilized our CapisciAMe software (1.3.76 version) [30] to construct and empower the first Italian dysarthric corpus aimed at AI research that exclusively comprises atypical speech utterances. Unlike other databases in the literature [31], our private speech collection, named CapisciAMe database, contains no audio data from individuals with typical (also known as normal) voices. Furthermore, we refrained from conducting any data augmentation operations on the collected data, distinguishing our approach from recent studies that create synthetic dysarthric data [9,32]. In recent years, the content of our database has been enriched both in terms of repetitions per keyword and in terms of the ASR dictionary size; nowadays, its total size and its inner organization outperform other Italian initiatives [33]. At the time of this writing, the CapisciAMe database was built owing to the collaboration of 208 Italian speakers with speech disabilities resulting from various neurological conditions including infant cerebral palsy (CP), stroke, and physical and traumatic brain injuries (TBIs), as well as neurodegenerative diseases, like Huntington's chorea. In total, our speech collection contains 65,282 unique speech samples, amounting to an overall recording time of 46.4 h. Each element within our labeled database represents a single atypical pronunciation of a keyword from our closed dictionary, produced by an anonymized individual with speech impairments. The corresponding waveform is sampled at 16 KHz and stored in a single-channel 16-bit PCM WAV file. The dataset content is not balanced due to variations in the number of samples collected by each speaker for the 79 distinct classes (or categories) within our ASR dictionary. These classes encompass a wide range of content, including numbers from zero to ten and voice commands for controlling smart home devices such as plugs, lamps, and televisions, as well as commands for music playback and vocal interaction with smartphones and computers. Furthermore, we meticulously reviewed the speech content to filter out background noise and unwanted audio components, ensuring the integrity of the atypical speech information for training deep learning architectures [34].

In this study, we used the CapisciAMe database to explore its synergy with a state-of-the-art (SOTA) ASR architecture, which is OpenAI's Whisper, and we specifically investigated the fine-tuning of three configurations of the same model with our corpus of Italian atypical speech. Whisper is currently trained in a fully supervised manner by exploiting a total amount of 680K hours of labeled speech data from multiple sources, including a block of multilingual data. The model is based on an encoder–decoder Transformer, which is fed by 80-channel log-Mel spectrograms obtained by the input speech waveform. Notably, the encoder is formed by two convolution layers with a kernel size of 3, followed by a sinusoidal positional encoding and a stacked set of Transformer blocks, whereas the decoder uses the learned positional embeddings and the same number of Transformer blocks from the encoder [7]. It is depicted in Figure 1. In this way, Whisper can map sequences of speech frames into sequences of characters. Diverse Whisper variants are available today with variations in the number of layers and attention heads. Specifically, in the present study, we employed the following configurations:

- Small, having approximately 244 million of parameters;
- Base, having approximately 74 million of parameters;
- Tiny, having approximately 39 million of parameters.

The proposed experimental evaluation conducted on such Whisper variants is detailed in the next section.

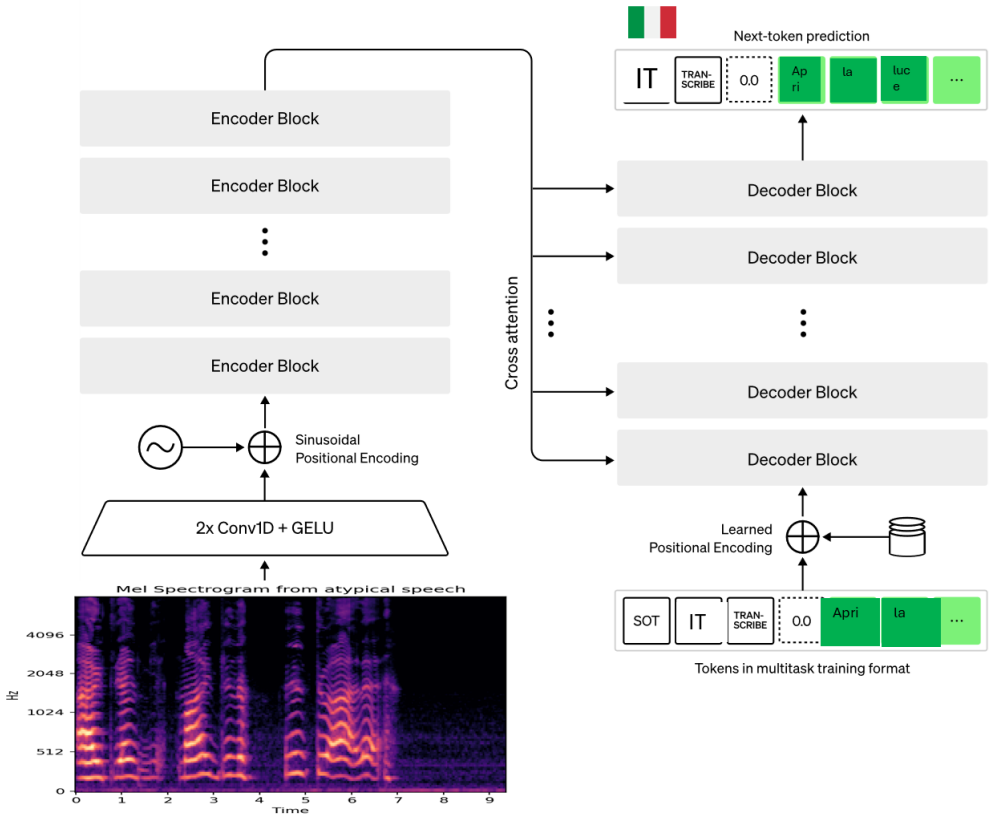

Figure 1. Structure of the Whisper model.

4. Experimental Evaluation

The following experiments had distinct objectives. First, with the collaborative effort of sixteen individuals having speech and motor disabilities, we studied the performance of the aforementioned variants of OpenAI's Whisper ASR architecture by considering the Word Recognition Accuracy (WRA) as the main performance indicator. This is a common metric for isolated word recognition tasks, and it expresses the percentage of the number of words the ASR models correctly identify to the number of words attempted, i.e., the size of our testing dataset. We employed this metric because we treated the problem of recognizing single speech commands as a multi-class classification task.

Next, we investigated the possibility of running the inference tasks (of the three models) on edge computing nodes, in particular single board computers, to the end of providing insights toward the development of a local Voice User Interface (VUI) for disordered speech, which is currently of great demand in the assistive technology field.

4.1. Speech Database Organization and Participants

The entire content of the Italian CapisciAMe database was exploited in the experiments. Our collection encompasses a total of 65,282 single speech recordings, corresponding to an overall duration time of 46.4 h. These speech contributions were created by 208 anonymized Italian speakers who had a condition causing a speech disorder. Among the total population, a subgroup of sixteen participants, whose details are reported in Table 2, was selected. These individuals were affected by motor disorders and varying severity levels of dysarthria, ranging from mild to severe. The considered speakers accounted for a large number of samples within our database, so, for each of them, 10 percent of the voice recordings was randomly selected to compose a testing dataset used to measure the performance of the proposed speech recognition solution. Moreover, the testing dataset was enriched with waveforms containing background noise and no human voice signal. As a result, the built datasets did not share any common elements: specifically, the training dataset included a total of 60,756 samples (recording time: 43.2 h) authored

by 208 distinct speakers, whereas the testing dataset comprised a total of 4526 samples (recording time: 3.2 h) created by the above selected participants.

Table 2. Sixteen Italian speakers engaged in the experiments.

Speaker	Gender	Age	Speech Disorder Cause	Speech Disorder Degree
IT01	M	39	CP	Moderate
IT02	M	53	CP	Severe
IT03	M	65	Neurodegenerative illnes	Mild
IT04	F	47	CP	Severe
IT05	M	41	CP	Moderate
IT06	M	72	Cerebropathy	Moderate
IT07	F	66	CP	Severe
IT08	F	34	CP	Mild
IT09	M	55	TBI	Mild
IT10	M	29	CP	Moderate
IT11	F	43	Cerebropathy	Mild
IT12	M	46	TBI	Severe
IT13	M	36	TBI	Moderate
IT14	M	40	CP	Severe
IT15	M	26	CP	Moderate
IT16	F	35	Aphasia	Moderate

4.2. Speech Models' Fine-Tuning

With the training dataset content, we fine-tuned three variants of OpenAI's Whisper model, i.e., small, base, and tiny. The fine-tuning process was resource-intensive and carried out on a single computation node within the Artemis High-Performance Computing (HPC) cluster at the University of Sydney. Our hardware setup included a machine equipped with 32GB of RAM and a single NVIDIA V100 GPU. We leveraged Python 3.7 as the programming language and PyTorch 1.3 as the deep learning framework, alongside NVIDIA CUDA 10.2 libraries. The Transformers library (version 4.26.1) from the HuggingFace platform facilitated the fine-tuning operations. Specifically, all the Whisper variants were trained for three epochs by using the same hyperparameters configuration reported in Table 3. Following fine-tuning, the relative ASR model checkpoints were generated and transferred on our local workstation to conduct inference experiments by considering the same testing dataset.

Table 3. Hyperparameters configuration.

Hyperparameter	Setup
Batch size	16
Warm up steps	500
Learning rate	0.00001
Number of epochs	3
Optimizer	AdamW
16-bit precision training	True

4.3. Word Recognition Accuracy Results

Table 4 contains the experimental results in terms of the percentage of Word Recognition Accuracy (WRA). It is defined by the following formula:

$$\text{WRA} = \frac{\text{Correct predictions}}{\text{All predictions}} \times 100 \quad (1)$$

By considering three variants of OpenAI's Whisper and three grades of speech impediment (mild, moderate, severe), we obtained the following results:

- A WRA of 95.9% by using the small variant;
- A WRA of 92.6% by using the base variant;
- A WRA of 90.1% by using the tiny variant.

Table 4. Performance of the three ASR models in terms of WRA (%) across the selected speakers.

		ASR Models			Testing Dataset	
Speech Disorder Severity	**Speaker**	**Small**	**Base**	**Tiny**	**Total Examples**	**Distinct Classes**
Severe	IT02	94.6	91.1	87.7	193	53
	IT04	*91.4*	*74.2*	*81.7*	83	13
	IT07	*88.1*	*89.3*	*85.7*	74	13
	IT12	*87.2*	*83.7*	*84.9*	76	13
	IT14	*90.4*	*87.8*	*74.7*	219	55
Average WRA		**90.3**	**85.2**	**82.9**		
Moderate	IT01	*97.8*	*93.0*	*92.5*	1343	79
	IT05	*92.0*	*87.0*	*89.0*	90	19
	IT06	*91.5*	*91.5*	*83.1*	108	38
	IT10	*94.4*	*91.0*	*88.8*	168	42
	IT13	*82.3*	*92.4*	*84.8*	69	13
	IT15	*96.2*	*92.4*	*89.2*	278	66
	IT16	*97.4*	*95.0*	*93.0*	332	43
Average WRA		**93.1**	**91.8**	**88.6**		
Mild	IT03	*91.0*	*92.3*	*88.5*	68	13
	IT08	*99.0*	*96.3*	*94.4*	1123	76
	IT09	*95.4*	*92.4*	*87.8*	227	46
	IT11	*89.4*	*88.2*	*80.0*	75	13
Average WRA		**93.7**	**92.3**	**87.7**		

To delve deeper, we analyzed individual testing datasets for each trial participant. These datasets contained varying numbers of keywords and diverse amounts of speech recordings per keyword from our dictionary. Notably, across all sixteen participants, the utilization of Whisper's small variant consistently led to improved WRA compared to the other configurations. The impact was most pronounced in cases of severe speech impediment, where the small variant exhibited approximately a 5% WRA improvement over the base variant. Comparing the base and tiny Whisper configurations, we observed a more modest increase. Similar trends were evident for moderate and mild speech disorders, as depicted in the bar graph shown in Figure 2. However, it is worth noting that the base model faced challenges when dealing with severe dysarthria impairment, particularly in instances like those of speakers IT04 and IT12. The reasons behind this abnormal performance remain elusive. We speculate that the extremely acute dysarthria exhibited by

these speakers, coupled with the scarcity of their speech samples (less than 100 examples across 13 distinct classes), led to significant divergence from the speech features present in our CapisciAMe corpus of atypical voice samples. At the same time, we note that a relationship between the number of the collected utterances and the WRA levels was difficult to establish.

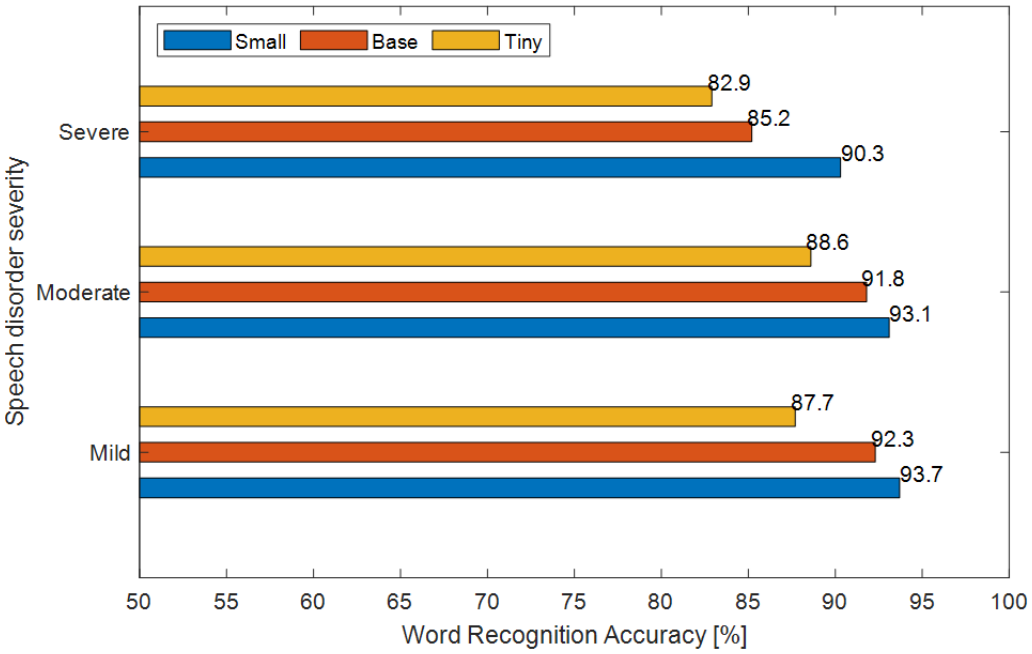

Figure 2. Performance of the three ASR models in terms of WRA (%) across diverse speech disorder severity levels.

As shown in this Section, the Whisper models, especially those incorporating the small variant, demonstrated promising results in handling speech impediments, but further investigation is needed to understand the nuances of the atypical speech and improve the models' robustness.

In the framework of disordered speech recognition, the quantitative results obtained in the current experiments outperformed previous investigations, particularly those that did not employ fine-tuning techniques [35]. In [36], a Word Error Rate (WER) of approximately 15% was measured using a multi-layer perceptron model trained (from scratch, i.e, no fine-tuning) using only a subset of the CapisciAMe speech collection.

Different results can be found in [37], where the conducted experiments focused on a limited number of speakers with speech impairments due to infant cerebral palsy. Further details are summarized in Table 5. However, we caution against relying solely on numerical accuracy values for direct comparisons. Researchers have employed diverse corpora and methodologies to evaluate ASR solutions, leading to variations in results. Factors such as the structure of the speech database, the chosen language, input modality, and varying levels of speech disability all significantly influence outcomes.

Table 5. Comparison between our study and previous articles in the literature.

Reference	Trial Participants	Classes in ASR Dictionary	Best Accuracy Result	Comments
[30]	3	13	94.4% WRA	Two-layer CNN model
[35]	6	13	95.6% WRA	Two-layer CNN model

Table 5. *Cont.*

Reference	Trial Participants	Classes in ASR Dictionary	Best Accuracy Result	Comments
[36]	16	13	84.4% WRA	Four-layer CNN model
[37]	2	11	92.9% WRA	Two-layer CNN model
This study	16	79	95.9% WRA	Fine-tuning Whisper

4.4. Design of a Voice User Interface on the Edge

With the expression "Voice User Interface" (VUI), we refer to a type of human–computer interaction that enables spoken communication with electronic devices. VUIs utilize ASR services to understand spoken commands and typically employs text-to-speech interfaces to deliver responses. A voice command device is one controlled through a VUI. Notable examples include smart speakers, which leverage virtual assistants to facilitate hands-free interactions in intelligent environments like smart homes.

As described below, we extended this mode of interaction to users with speech disorders by exploiting the ASR solutions previously analyzed. Specifically, in order to achieve local ASR inference without internet dependency, we deployed the Whisper variants (small, based, tiny) on edge computing nodes—specifically, single-board computers. Our chosen platform for this investigation was one single Raspberry Pi 5 board. The key steps in our approach involved porting the fine-tuned ASR models to the edge environment using the "Whisper.cpp" package (https://github.com/ggerganov/whisper.cpp, accessed on 27 January 2024), which provides a solution for executing inference tasks in the C++ language. As recommended by the creator, we converted the speech models into the ggml format. Subsequently, we evaluated the performance of our models on the Raspberry Pi, focusing on overall inference times as the primary metric of interest. For each of the three fine-tuned ASR models, the proposed investigation involved a testing dataset comprising 50 randomly selected atypical speech samples previously used in word recognition accuracy experiments. We analyzed both the total inference time required for the atypical speech transcription and the associated hardware resources, specifically RAM and disk usage. Our findings are summarized in Table 6. We observe the following:

- The small variant, which has more parameters in its architecture compared to the other models, demanded substantial resources of our embedded system. According to the overall WRA results, it can work well in ASR, but its utilization resulted in a mean inference time of approximately 8.5 s, which is impractical for real-time applications.
- The tiny model exhibited an interesting behavior. It achieved an average inference time of 1.2 s (acceptable for several application scenarios) and utilized around 400MB of RAM. However, its recognition accuracy currently suffers in cases of severe dysarthria.
- The base variant fell in between the small and tiny architectures. It yielded a mean inference time of 2.6 s and utilized more embedded system resources in comparison to the tiny variant.

As a consequence, identifying the more appropriate Whisper configuration for handling atypical speech patterns remains a challenging task to accomplish. Despite this, we believe that the creation of a specific VUI running on edge computing nodes may act as an enabler for the development of customized applications for users with disordered voices. For instance, the embedded computer can work as a voice-input voice-output communication aid, helping its user to convert personal utterances in more complex sentences spoken aloud by a computer-generated voice. In this way, the speaker with dysarthria can rely on a set of keywords to express more clearly their or her personal needs, in order to facilitate the spoken interaction with caregivers and collaborators. A different scenario may

be concerned with the domain of the smart home automation, wherein the creation of alternative ways to have an interaction with virtual assistants' services is of great importance nowadays [38].

Table 6. Performance of the three ASR models on a Raspberry Pi 5 device.

	Inference Times [ms]		Hardware Resources [MB]	
ASR Models	**Mean**	**Std Dev**	**Disk Usage**	**RAM Usage**
Small	8575	93	476	1024
Base	2604	94	144	500
Tiny	1171	30	76	390

5. Conclusions

This study investigated the potential impact of a state-of-the-art ASR model on impaired speech recognition, and, specifically, we focused our attention on OpenAI's Whisper. The conducted experimental evaluation highlighted the effectiveness of our proposed methodology, wherein we fine-tuned three distinct model variants (small, base, and tiny) by using our private database of Italian atypical speech. This corpus comprises over 65 K voice samples contributed by 208 individuals with speech impairments, including conditions like dysarthria. Overall, we obtained positive results in terms of word recognition accuracy, particularly in the small configuration, achieving a score of 95.9% (resulting in a word error rate of approximately 4%) when evaluated against our closed ASR dictionary containing 79 unique keywords. Furthermore, we explored the deployment of these trained models on edge computing nodes, envisioning the creation of customized voice user interfaces capable of interpreting disordered speech commands across diverse application contexts. In this context, it is evident that the tiny configuration model exhibits superior performance in terms of overall computation times in comparison to the other examined configurations.

The present work holds significant implications for assistive technologies; however, we acknowledge important limitations. The number of trial participants with speech impediments restricts the generalizability of our approach to various forms of speech disabilities. Additionally, our methodology is specific to the Italian language, rendering it unsuitable for direct application to other languages. To address these challenges, our future studies will extend investigations to different languages, including English, and explore additional variants of the Whisper model to enhance voice user interfaces for individuals with impaired speech across various application scenarios.

Author Contributions: Conceptualization, D.M.; methodology, D.M.; software, D.M.; formal analysis, D.M.; investigation, D.M.; data curation, D.M.; writing—original draft preparation, responses to reviewers' comments, D.M.; supervision, M.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available on reasonable request from the corresponding author. The data are not publicly available due to privacy restrictions.

Acknowledgments: The authors express gratitude to the Sydney Informatics Hub and the University of Sydney's high-performance computing cluster, Artemis, for furnishing the computational resources essential to the research findings detailed in this paper. Davide Mulfari is an affiliate researcher at the University of Sydney. Currently, he is attending the National Ph.D. in Artificial Intelligence, XXXVII cycle, a course in health and life sciences, organized by Università Campus Bio-Medico di Roma.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Stewart, C.; Riedel, K. Managing Speech and Language Deficits after Stroke. In *Stroke Rehabilitation*, 4th ed.; Gillen, G., Ed.; Mosby: London, UK, 2016; pp. 673–689. [CrossRef]
2. De Russis, L.; Corno, F. On the impact of dysarthric speech on contemporary ASR cloud platforms. *J. Reliab. Intell. Environ.* **2019**, *5*, 163–172. [CrossRef]
3. Ballati, F.; Corno, F.; Russis, L.D. *Assessing Virtual Assistant Capabilities with Italian Dysarthric Speech*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 93–101. [CrossRef]
4. Jaddoh, A.; Loizides, F.; Rana, O.; Syed, Y.A. Interacting with Smart Virtual Assistants for Individuals with Dysarthria: A Comparative Study on Usability and User Preferences. *Appl. Sci.* **2024**, *14*, 1409. [CrossRef]
5. Lin, Y.; Wang, L.; Dang, J.; Li, S.; Ding, C. Disordered speech recognition considering low resources and abnormal articulation. *Speech Commun.* **2023**, *155*, 103002. [CrossRef]
6. Shor, J.; Emanuel, D.; Lang, O.; Tuval, O.; Brenner, M.; Cattiau, J.; Vieira, F.; McNally, M.; Charbonneau, T.; Nollstadt, M.; et al. Personalizing ASR for Dysarthric and Accented Speech with Limited Data. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 784–788. [CrossRef]
7. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518. [CrossRef]
8. Shahamiri, S.R. Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 852–861. [CrossRef]
9. Almadhor, A.; Irfan, R.; Gao, J.; Saleem, N.; Tayyab Rauf, H.; Kadry, S. E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition. *Expert Syst. Appl.* **2023**, *222*, 119797. [CrossRef]
10. Enríquez, J.; Soria Morillo, L.M.; García-García, J.A.; Álvarez-García, J.A. Two decades of assistive technologies to empower people with disability: A systematic mapping study. In *Disability and Rehabilitation: Assistive Technology*; Taylor & Francis Group Limited: Oxford, UK, 2023; pp. 1–18. [CrossRef]
11. Qian, Z.; Xiao, K. A Survey of Automatic Speech Recognition for Dysarthric Speech. *Electronics* **2023**, *12*, 4278. [CrossRef]
12. Bharti, K.; Das, P.K. A Survey on ASR Systems for Dysarthric Speech. In Proceedings of the 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), Delhi, India, 9–10 December 2022; pp. 1–6. [CrossRef]
13. Hawley, M.S.; Enderby, P.; Green, P.; Cunningham, S.; Palmer, R. Development of a Voice-Input Voice-Output Communication Aid (VIVOCA) for People with Severe Dysarthria. In *Computers Helping People with Special Needs*; Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 882–885. [CrossRef]
14. Cunningham, S.; Green, P.; Christensen, H.; Atria, J.; Coy, A.; Malavasi, M.; Desideri, L.; Rudzicz, F. *Harnessing the Power of Technology to Improve Lives*; IOS Press: Amsterdam, The Netherlands, 2017; Volume 242, pp. 322–329.
15. Malavasi, M.; Turri, E.; Atria, J.; Christensen, H.; Marxer, R.; Desideri, L.; Coy, A.; Tamburini, F.; Green, P. An Innovative Speech-Based User Interface for Smarthomes and IoT Solutions to Help People with Speech and Motor Disabilities. *Stud. Health Technol. Inform.* **2017**, *242*, 306. [PubMed]
16. Donati, M.; Bechini, A.; D'Anna, C.; Fattori, B.; Marini, M.; Olivelli, M.; Pelagatti, S.; Ricci, G.; Schirinzi, E.; Siciliano, G.; et al. A Clinical Tool for Prognosis and Speech Rehabilitation in Dysarthric Patients: The DESIRE Project. In *Applications in Electronics Pervading Industry, Environment and Society*; Berta, R., De Gloria, A., Eds.; Springer: Cham, Switzerland, 2023; pp. 380–385. [CrossRef]
17. Graham, C.; Roll, N. Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Lett.* **2024**, *4*, 025206. [CrossRef]
18. Barcovschi, A.; Jain, R.; Corcoran, P. A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving the child speech recognition. In Proceedings of the 2023 International Conference on Speech Technology and Human–Computer Dialogue (SpeD), Bucharest, Romania, 25–27 October 2023; pp. 42–47. [CrossRef]
19. Rathod, S.; Charola, M.; Patil, H.A. Transfer Learning Using Whisper for Dysarthric Automatic Speech Recognition. In *International Conference on Speech and Computer*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 579–589. [CrossRef]
20. Sanguedolce, G.; Naylor, P.A.; Geranmayeh, F. Uncovering the Potential for a Weakly Supervised End-to-End Model in Recognising Speech from Patient with Post-Stroke Aphasia. In Proceedings of the 5th Clinical Natural Language Processing Workshop, Toronto, Canada, 14 July 2023; pp. 182–190. [CrossRef]
21. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460. [CrossRef]
22. Matsushima, T. Dutch Dysarthric Speech Recognition: Applying Self-Supervised Learning to Overcome the Data Scarcity Issue. Ph.D. Thesis, University of Groningen, Groningen, The Netherlands, 2022.
23. Baskar, M.K.; Herzig, T.; Nguyen, D.; Diez, M.; Polzehl, T.; Burget, L.; Černockỳ, J. Speaker adaptation for Wav2vec2 based dysarthric ASR. *arXiv* **2022**, arXiv:2204.00770. [CrossRef].
24. Violeta, L.P.; Huang, W.C.; Toda, T. Investigating Self-supervised Pretraining Frameworks for Pathological Speech Recognition. *arXiv* **2022**, arXiv:2203.15431. [CrossRef].
25. Hernandez, A.; Pérez-Toro, P.A.; Nöth, E.; Orozco-Arroyave, J.R.; Maier, A.; Yang, S.H. Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition. *arXiv* **2022**, arXiv:2204.01670. [CrossRef]

26. Wang, P.; Van Hamme, H. Benefits of pre-trained mono-and cross-lingual speech representations for spoken language understanding of Dutch dysarthric speech. *EURASIP J. Audio Speech Music Process.* **2023**, *2023*, 15. [CrossRef]
27. Hu, S.; Xie, X.; Jin, Z.; Geng, M.; Wang, Y.; Cui, M.; Deng, J.; Liu, X.; Meng, H. Exploring Self-Supervised Pre-Trained ASR Models for Dysarthric and Elderly Speech Recognition. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [CrossRef]
28. Torre, I.G.; Romero, M.; Álvarez, A. Improving Aphasic Speech Recognition by Using Novel Semi-Supervised Learning Methods on AphasiaBank for English and Spanish. *Appl. Sci.* **2021**, *11*, 8872. [CrossRef]
29. Young, V.; Mihailidis, A. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assist. Technol.* **2010**, *22*, 99–112. [CrossRef]
30. Mulfari, D.; Meoni, G.; Marini, M.; Fanucci, L. Machine learning assistive application for users with speech disorders. *Appl. Soft Comput.* **2021**, *103*, 107147. [CrossRef]
31. Kim, H.; Hasegawa-Johnson, M.; Perlman, A.; Gunderson, J.; Huang, T.S.; Watkin, K.; Frame, S. Dysarthric speech database for universal access research. In Proceedings of the Interspeech 2008, Brisbane, Australia, 22–26 September 2008; pp. 1741–1744. [CrossRef]
32. Shahamiri, S.R.; Lal, V.; Shah, D. Dysarthric Speech Transformer: A Sequence-to-Sequence Dysarthric Speech Recognition System. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 3407–3416. [CrossRef] [PubMed]
33. Turrisi, R.; Braccia, A.; Emanuele, M.; Giulietti, S.; Pugliatti, M.; Sensi, M.; Fadiga, L.; Badino, L. EasyCall corpus: A dysarthric speech dataset. *arXiv* **2021**, arXiv:2104.02542. [CrossRef] .
34. Mulfari, D.; Campobello, G.; Gugliandolo, G.; Celesti, A.; Villari, M.; Donato, N. Comparison of Noise Reduction Techniques for Dysarthric Speech Recognition. In Proceedings of the 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Messina, Italy, 22–24 June 2022; pp. 1–6. [CrossRef]
35. Mulfari, D.; La Placa, D.; Rovito, C.; Celesti, A.; Villari, M. Deep learning applications in telerehabilitation speech therapy scenarios. *Comput. Biol. Med.* **2022**, *148*, 105864. [CrossRef] [PubMed]
36. Mulfari, D.; Carnevale, L.; Villari, M. Toward a lightweight ASR solution for atypical speech on the edge. *Future Gener. Comput. Syst.* **2023**, *149*, 455–463. [CrossRef]
37. Mulfari, D.; Carnevale, L.; Galletta, A.; Villari, M. Edge Computing Solutions Supporting Voice Recognition Services for Speakers with Dysarthria. In Proceedings of the 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW), Bangalore, India, 1–4 May 231–236. [CrossRef]
38. Jaddoh, A.; Loizides, F.; Lee, J.; Rana, O. An interaction framework for designing systems for virtual home assistants and people with dysarthria. In *Universal Access in the Information Society*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 1–13. [CrossRef]