ID N. AIDR01/16913

UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

UNIVERSITÀ DEGLI STUDI DI MESSINA

DEPARTMENT OF MATHEMATICAL AND COMPUTER SCIENCES, PHYSICAL SCIENCES AND EARTH SCIENCES

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVII Cycle

Exploring Artificial Intelligence Technologies for Automatic Speech Recognition in Voice Disorders

Supervisor
Prof. Massimo Villari

Candidate
Davide Mulfari

March, 2025

"Nothing about us, without us"

Abstract

This thesis focuses on the application of Artificial Intelligence (AI)-driven techniques to support Automatic Speech Recognition (ASR) services for individuals with speech impairments, such as dysarthria. These conditions often result in poor speech intelligibility and are frequently accompanied by severe motor disabilities. It is estimated that over 22 million people in Europe (5% of the population) are affected by such speech disorders, which can manifest from childhood, as in the case of cerebral palsy, or due to neurodegenerative and progressive diseases like Parkinson's, amyotrophic lateral sclerosis, spinal muscular atrophy, and multiple sclerosis. These individuals face significant barriers in interpersonal communication due to articulation problems that produce extremely variable speech. These impairments impose profound limitations on social participation and independence in daily activities.

While contemporary ASR tools integrated into voice assistants excel with standard speech, their performance deteriorates significantly when faced with impaired speech patterns, particularly moderate to severe voice disorders. This creates a paradox: technologies that could be crucial in improving the lives of people with the disabilities instead become additional barriers.

To address this, the present work proposes a technological ecosystem called Capisci-AMe, designed for disordered speech recognition with a focus on Italian as the primary language. The research is centered on isolated word recognition tasks using speaker-dependent approaches and leverages deep learning techniques, including state-of-the-art ASR models based on encoder-decoder architectures. These models are fine-tuned with our private corpus of Italian impaired speech, enabling progress toward recognizing short sentences as combinations of individual words. The proposed ecosystem encompasses three interrelated pillars, each a crucial aspect of this research:

- Disordered Speech Collection: Given the scarcity of dysarthric speech corpora, especially in Italian, a significant effort has been devoted to collecting voice samples from individuals with speech disorders. This work has resulted in the development of the first Italian atypical speech corpus for AI-based research. To support this effort, novel IoT-based assistive technologies have been introduced to streamline speech acquisition, alongside methodologies for enhancing impaired speech signals.

- Deep Learning Architectures: The study employs sequence-to-sequence frameworks, leveraging state-of-the-art encoder-decoder architectures such as Wav2Vec2 (by Meta AI) and Whisper (by OpenAI). These models, based on transformer and pre-trained on extensive multilingual standard speech datasets, were fine-tuned on our private corpus.

This fine-tuning process is critical to our work, as it enables accurate recognition of single voice commands and short sentences (as combinations of isolated words) spoken by Italian individuals with atypical speech and dysarthria.

- ASR Services and Application Prototypes: A cloud-based speech-to-text transcription service, powered by the ASR engine, has been developed as part of the CapisciAMe ecosystem. This platform facilitates seamless integration of speech recognition capabilities into custom applications, empowering software developers and fostering interdisciplinary studies and applications that support individuals with speech impairments.

This research demonstrates the feasibility of creating tailored ASR systems for disordered speech by addressing challenges in data scarcity, model optimization, and application accessibility. The ecosystem achieves significant improvements in recognizing impaired speech in Italian, laying the groundwork for further development of inclusive communication technologies that enhance the independence and social participation of individuals with speech impairments.

Contents

List of Figures

List of Tables

Earlier Publications

This thesis marks the culmination of my doctoral studies in Artificial Intelligence. The findings from this research have been shared through various international conference proceedings and journal publications, as listed below.

Journals:

- Mulfari, D., & Villari, M. (2024). A Voice User Interface on the Edge for People with Speech Impairments. Electronics, 13(7), 1389.
  DOI: 10.3390/electronics13071389

- Pucci, E., Piro, L., Possaghi, I., Mulfari, D., & Matera, M. (2024). Co-designing the integration of voice-based conversational AI and web augmentation to amplify web inclusivity. Scientific Reports, 14(1), 16162.
  DOI: 10.1038/s41598-024-66725-3

- Mulfari, D., Carnevale, L., & Villari, M. (2023). Toward a lightweight ASR solution for atypical speech on the edge. Future Generation Computer Systems, 149, 455-463.
  DOI: 10.1016/j.future.2023.08.002

- Mulfari, D., La Placa, D., Rovito, C., Celesti, A., & Villari, M. (2022). Deep learning applications in telerehabilitation speech therapy scenarios. Computers in Biology and Medicine, 148, 105864.
  DOI: 10.1016/j.compbiomed.2022.105864

International conference proceedings:

- Mulfari, D., Carnevale & Villari, M. (2024, June). Sequence-to-Sequence Models in Italian Atypical Speech Recognition. In 2024 IEEE Symposium on Computers and Communications (ISCC) (pp. 1-6). IEEE.
  DOI: 10.1109/ISCC61673.2024.10733600

- Mulfari, D., Carnevale, L., Galletta, A., & Villari, M. (2023, May). Edge Computing Solutions Supporting Voice Recognition Services for Speakers with Dysarthria. In 2023

IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW) (pp. 231-236). IEEE.
DOI: 10.1109/CCGridW59191.2023.00047

- Mulfari, D., Campobello, G., Gugliandolo, G., Celesti, A., Villari, M., & Donato, N. (2022, June). Comparison of Noise Reduction Techniques for Dysarthric Speech Recognition. In 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA) (pp. 1-6). IEEE.
DOI: 10.1109/MeMeA54994.2022.9856486

- Mulfari, D., Celesti, A., & Villari, M. (2022, May). Exploring AI-based Speaker Dependent Methods in Dysarthric Speech Recognition. In 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid) (pp. 958-964). IEEE.
DOI: 10.1109/CCGrid54584.2022.00117

Chapter 1

Introduction

Speech plays a crucial role in human communication, allowing individuals to express thoughts, emotions, and needs effectively. However, this fundamental ability is significantly hindered for individuals with speech impairments, such as dysarthria. Dysarthria, a prevalent neurological speech disorder, affects approximately 22 million people across Europe, representing 5% of the population. It arises from weaknesses in the muscles responsible for speech, miscoordination or inaccuracy in articulatory movements, and irregularities in tone, steadiness, or speed. As a result, the produced speech is atypical and characterized by prosodic features such as monoloudness, monopitch, and impaired ranges of fundamental frequency, rate, and vocal intensity [1].

This speech disorder frequently accompanies severe motor impairments, as observed in individuals with cerebral palsy or neurodegenerative diseases like amyotrophic lateral sclerosis, spinal muscular atrophy, multiple sclerosis, and Parkinson's disease. These conditions result in significantly limited motor skills, coupled with irregular speech patterns and reduced speech clarity. Together, these challenges severely hinder effective interpersonal communication, imposing substantial limitations on social participation and independence in daily activities. For many, these barriers can lead to social exclusion, profoundly affecting the quality of life for millions of individuals worldwide.

In this context, speech-controlled interfaces represent a transformative opportunity in assistive technologies, particularly for individuals with combined speech and motor impairments. Traditional methods of interaction, such as touchscreens, mice, or keyboards, are often inaccessible or cumbersome for those with reduced motor skills. Speech-controlled systems, when designed to accommodate atypical speech patterns, provide a hands-free, intuitive means of communication and control. These systems empower users to interact with digital devices and environments, ranging from smartphones to home automation technologies, offering greater autonomy and access to essential services.

For individuals with speech impairments, such interfaces can act as a bridge to regain independence and foster more inclusive interactions. By accurately recognizing their speech, users can issue commands, dictate text, and navigate digital environments, significantly enhancing their quality of life. However, the efficacy of these systems depends on their ability to adapt to the unique characteristics of impaired speech. Current speech recognition technologies often struggle to handle speech disorders effectively, particularly in moderate to severe cases, where recognition performance remains inadequate [2]. Overcoming these challenges demands innovative, tailored solutions to ensure that speech-controlled technologies are not only functional but also inclusive and empowering for this population.

1.1 Motivation

Automatic Speech Recognition (ASR) is a technology designed to map speech waveforms into appropriate sequences of characters that can be processed by machines. In the field of human-computer interaction, ASR holds crucial importance as it powers automated transcription services, enabling Artificial Intelligence (AI)-based conversational agents and widely used virtual assistants' services that operate through voice commands. These systems leverage Voice User Interfaces (VUIs) to process and respond to human speech, offering natural and intuitive interactions. The integration of ASR technology into various applications has made it a cornerstone of modern digital ecosystems. From smart home automation to hands-free interactions with computers and connected devices, ASR has become pervasive in a wide range of scenarios. However, while these advancements enhance convenience for many, they often exclude individuals with speech impairments, such as dysarthria.

For individuals with dysarthria, communicating effectively is a significant challenge, particularly with unfamiliar human communication partners and machines. Dysarthria is frequently associated with motor disabilities, such as spastic quadriplegia, compounding the difficulties these individuals face in everyday activities. Beyond communication barriers, their interaction with the surrounding environment is also limited, such as the ability to perform simple actions (for example turning on the light by pressing a common switch). Therefore, ASR technologies have the potential to be life-changing for this population. By enabling computers and portable digital devices to serve as interaction mediums [3], ASR can empower individuals with speech impairments to communicate with others and interact with digital systems. Despite this potential, current ASR systems perform poorly when recognizing disordered speech, particularly due to the unique challenges posed by non-standard speech patterns caused by voice disorders [4].

Dysarthria significantly affects the articulation of phonemes, particularly among individ-

uals with severe disabilities. Produced phonemes often lack clarity, resulting in imprecise pitch, pauses, and disruptions in the production of consonants and vowels. These inaccuracies obscure the distinct auditory features that ASR systems depend on to differentiate between phonemes. Moreover, the variability in the effects of disabilities introduces a wide range of speech variations among individuals with dysarthria, far exceeding those observed in typical speech. As a result, standard ASR approaches struggle to accurately map dysarthric speech to corresponding transcription. These systems are inadequate to address challenges such as incorrect tempo, irregular phonation, and inconsistencies in formant shifting, leading to poor recognition performance. Research [5] has shown that conventional ASR systems fail to handle these complexities effectively, with their performance deteriorating as the severity of dysarthria increases. A comprehensive review of ASR technologies reveals that while traditional systems may perform reasonably well with mild dysarthria, their accuracy declines sharply for moderate to severe cases. Studies consistently report high word error rates (WERs) when standard ASRs process dysarthric speech, highlighting their inability to accurately recognize non-standard speech patterns. This limitation underscores the need for innovative approaches tailored to the unique characteristics of dysarthric speech to improve recognition accuracy and usability for affected individuals [6, 7]. Furthermore, a major challenge in developing accurate ASR systems for dysarthric speech lies in the substantial variability of dysarthric speech patterns. This variability demands a large volume of dysarthric speech data to train ASR acoustic models effectively. However, the availability of such data is extremely limited, as collecting speech samples from individuals with dysarthria is a complex and resource-intensive process. Also, the inherent difficulties faced by individuals with dysarthria in producing clear and consistent sounds add another layer of complexity. These challenges make precise phoneme labeling for dysarthric speech particularly arduous, as the atypical nature of the speech introduces inaccuracies and inconsistencies that are difficult to annotate accurately. This lack of robust data and reliable labeling significantly hampers the development of ASR systems capable of handling disordered speech patterns [8, 9]. It underscores the current limitations of the today's technologies in recognizing atypical speech and reveal a bias in the multilingual databases used to train current ASR systems, where non-standard speech patterns are underrepresented.

In recent years, automatic disordered speech recognition has attracted significant attention from both industry and academic research communities. While global efforts and major research projects have predominantly focused on the American language [10], there is currently no dedicated ASR solution for Italian-speaking individuals with disordered speech. To address this gap, this dissertation explores the application of AI techniques, particularly those based on deep learning, to develop services for the automated recognition of impaired

speech in Italian. The proposed research does not aim to solve the multiple challenges of automatic speech recognition in the presence of speech impairments and for connected (and continuous) speech in its complexity.

1.2 Contributions

The proposed research falls within the domain of digital assistive technologies for persons with disabilities. In this context, it is well-documented that users with disabilities often adapt and optimize their skills to access and interact with digital devices [11]. For instance, many individuals with neuromotor disabilities who are unable to use a traditional hand-driven mouse employ assistive software components to replicate its functionality. A common example is the use of a limited set of keys on the numerical keypad to perform all essential mouse actions, demonstrating how a small number of inputs can effectively enable full control over a computer.

We seek to map this principle into the field of speech recognition. Specifically, we emphasize the importance of developing an ASR system capable of recognizing a limited number of isolated words and short, meaningful sentences pronounced by speakers with disordered speech. Building upon this concept, we propose the design and initial implementation of a digital ecosystem supporting ASR services in Italian, named CapisciAMe. This ecosystem is underpinned by a complex architecture comprising three interconnected pillars, which represent the primary contributions of our work:

1. **Disordered Speech Collection**

   The scarcity of dysarthric speech corpora, particularly for the Italian language, represents a significant barrier to progress in this domain. Therefore, a key part of our work involves acquiring voice samples from individuals with speech disorders to address this gap. This ongoing effort plays a pivotal role in empowering the first Italian atypical speech corpus for AI-based research. To achieve this, we introduced novel assistive software solutions, including the CapisciAMe mobile app, designed to facilitate speech data acquisition directly from end users. Additionally, we developed procedures for enhancing impaired speech signals. As a result, the CapisciAMe database has emerged as the most comprehensive dataset of voice samples from Italian-speaking individuals with speech disorders, to the best of our knowledge.

2. **Deep Learning Architectures**

   This research employs sequence-to-sequence frameworks, leveraging state-of-the-art (SOTA) encoder-decoder architectures such as Wav2Vec2 and Whisper. These models,

pre-trained on extensive multilingual standard speech datasets, were fine-tuned on the entire CapisciAMe corpus. This fine-tuning process is critical to our work, as it enables accurate recognition of single voice commands and short sentences (as combinations of isolated words) spoken by Italian individuals with atypical speech and dysarthria. Experimental evaluations have highlighted the effectiveness of the proposed approach, with an overall word error rate of 3.5% measured on a specific testing dataset.

3. **ASR Services and Application Prototypes**

   To translate our research into practical impact, we propose integrating the ASR system into VUIs. Deploying our speech recognizer as a cloud-based service is essential for enabling real-world applications that leverage its capabilities. This on-demand speech-to-text conversion framework forms a vital component of the CapisciAMe ecosystem, offering software developers an accessible platform for embedding VUI features into custom applications. This approach not only broadens the reach of our ASR solutions but also opens avenues for supporting diverse studies and applications across fields, e.g., human computer interaction.

By addressing these three pillars, our CapisciAMe project is an example of AI for Social Good. It aims to represent a significant step toward developing inclusive ASR technologies tailored to the needs of individuals with disordered speech.

1.3 Research question

In this thesis, the research questions addressed pertain to the pillars of our digital ecosystem: impaired speech acquisition and preprocessing, deep learning model training for ASR, and the development of speech recognition services (based on the cloud computing paradigm) and application prototypes.

Regarding the first topic, one of the crucial innovations of our work is the use of smartphones (with a dedicated application) to collect impaired speech samples.

This leads to the research question (RQ1): *How should a mobile voice collection system be designed?* This will be discussed in Chapter 4

Regarding the second topic, one of the contributions of our thesis is the application of state-of-the-art ASR models (based on encoder-decoder architectures) to disordered speech.

This leads to the research question (RQ2): *What are the benefits of such methodologies for impaired ASR?* This will be analyzed in Chapter 5.

Additionally, concerning the third pillar of our digital ecosystem, a shared research question (RQ3) can be formulated as follows: *What are the potential applications of voice user*

interfaces for atypical speech? This will be discussed in Chapter 6.

1.4 Thesis organization

This thesis is organized into seven chapters, each covering distinct aspects of our ongoing research in the field of automatic disordered speech recognition.

Chapter 2 addresses the concept of speech disorders, with a focus on dysarthria, and outlines the main challenges tackled in this dissertation. In particular, Section 2.5 introduces our digital ecosystem, CapisciAMe, which supports speech recognition services for Italian-speaking individuals with speech impairments.

Chapter 3 presents a literature review in the field of automatic disordered speech recognition. The analysis delves into deep neural network ASR approaches and recent advanced solutions leveraging sequence-to-sequence structures.

Chapter 4 details our methodology for disordered speech collection and describes the features of our private corpus of impaired speech, the CapisciAMe database. To the best of our knowledge, this database represents the richest and most comprehensive archive in Italian, including voice samples from individuals with dysarthria and other speech disorders.

Chapter 5 discusses our deep learning approaches to impaired speech recognition. Specifically, Section 5.2 explores the utilization of pre-trained state-of-the-art encoder-decoder architectures, namely Wav2Vec2 (by Meta AI) and Whisper (by OpenAI), which have been fine-tuned on the CapisciAMe database. Experimental evaluations highlight the effectiveness of the proposed approaches and are detailed in Section 5.5.

Chapter 6 focuses on the creation of cloud-based transcription services and application prototypes that leverage our speech recognition engine in various scenarios within the framework of digital assistive technologies.

Finally, Chapter 7 concludes the thesis, highlighting possible future research directions for our project.

Chapter 2

Speech disorders

This chapter introduces the concept of speech disorder, with a focus on dysarthria. We also discuss the key challenges in automatic disordered speech recognition addressed in our thesis.

2.1 Introduction

Speech is the primary means of conveying thoughts, emotions, and ideas to others. It requires the complex coordination of various body parts, including the head, neck, and chest, to ensure effective communication. A speech disorder is a health condition that hinders a person's ability to articulate words, often due to damage to the muscles, nerves, or vocal structures involved in speech production. As such, a speech disorder (or impairment) encompasses deviations from normal speech patterns and difficulties in producing the sounds necessary for interpersonal communication. These conditions can occur independently or be associated with other disorders, such as cerebral palsy or neurodegenerative diseases. Various factors contribute to speech disorders, including developmental issues, neurological conditions, and physical abnormalities. Notable examples of speech impairments include articulation disorders (leading to difficulty pronouncing specific sounds or words), phonological disorders (with challenges in understanding and using speech patterns), fluency disorders (such as stuttering or cluttering), and voice disorders that determine abnormal pitch, volume, or voice quality.

Within this spectrum, specific speech disorders include:

- Dysarthria: A common motor speech disorder that affects the motor mechanism of speech and manifests as dysfunctions within the respiratory, phonatory, and articulatory apparatus, leading to irregularities in speech intelligibility. This condition often results in slurred, slow, or difficult to understand speech and is typically associated with neurological injuries or conditions.

- Apraxia: A motor speech disorder involves difficulty in planning and programming the movements necessary for speech, despite having the physical ability to speak. It is not due to muscle weakness but rather a disruption in the brain's ability to send the correct signals to the speech muscles.

- Vocal cord paralysis: A condition where one or both vocal cords (also known as vocal folds) lose their ability to move. This can significantly impact speaking, breathing, and swallowing.

- Spasmodic dysphonia: A rare neurological disorder that affects the muscles of the larynx. This condition causes involuntary spasms in the vocal cords, leading to interruptions in speech and changes in voice quality.

- Orofacial myofunctional disorders: Conditions related to abnormal movement patterns of the face and mouth that can affect speech, eating, and breathing. These disorders can interfere with the normal growth and development of the muscles and bones of the face and mouth.

Recent reports indicate that approximately 7.5 million people in the United States [1] live with speech impairments, while over 22 million individuals in Europe (5% of the population) experience speech impairments like dysarthria. Beyond the practical challenges of daily life, these disorders can have profound psychological and social effects on individuals and their families [12].

2.2 Dysarthria

Within the complex framework of the impairments leading to disordered (atypical) speech, in this thesis we focus primarily on dysarthria. This neuromotor speech disorder manifests as disturbances in muscular control over the speech mechanism due to a damage of the central or peripheral nervous system. It designates problems in oral communication due to paralysis, weakness, or incoordination of the speech musculature [13], and, at the same time, it leads to abnormalities in the strength, speed, range, steadiness, tone, and accuracy of movements required for speech production [1]. Dysarthria exhibits significant variability both within and across speakers, impacting speech intelligibility. Notably, phoneme pronunciation differs markedly in moderate and acute cases, setting dysarthric speech apart from standard speech.

[1] https://www.asha.org/

The above speech disorders co-exist with severe motor impairments and can arise congenitally or be acquired at any age due to neurologic injury, disease, or disorder. Clinical literature recognizes several primary types of dysarthria [14]:

- Flaccid: it is caused by damage to the lower motor neurons (cranial and spinal nerves). Symptoms include weakness, breathy voice, hypernasality, nasal emissions, and short phrases.
- Spastic: it results from bilateral damage to the upper motor neurons. It is characterized by a strained or harsh voice quality, slow speech rate, pitch breaks, and hypertonia (increased muscle tone).
- Ataxic: it is due to damage to the cerebellum. Symptoms include irregular speech rhythm, distorted vowels, excess and equal stress, and lack of coordination in jaw, face, and tongue movements.
- Hypokinetic: it is associated with damage to the basal ganglia, often seen in Parkinson's disease. It features monopitch, reduced loudness, rapid or accelerated speech rate, and rigidity.
- Hyperkinetic: it is caused by damage to the basal ganglia, often linked to conditions like Huntington's disease. Symptoms include involuntary movements, voice tremor, sudden forced inspiration or expiration, and variable speech rate.
- Mixed: it involves a combination of two or more types of dysarthria, often seen in conditions like amyotrophic lateral sclerosis (ALS) or multiple sclerosis (MS). Symptoms are a mix of characteristics from the involved types, such as spastic and flaccid features.

Despite the prevalence of these speech disorders in various neurologic conditions, the incidence and prevalence of dysarthria remain elusive nowadays. Estimates vary based on lesion location, underlying condition nature, and assessment criteria. Cerebral palsy (CP), i.e., a group of disorders affecting movement and posture, caused by non-progressive damage to the developing brain (typically before birth), stands out as a significant cause of dysarthric speech, affecting 90% of children with CP [15]. Other common sources of atypical speech include stroke (impacting 20% to 30% of survivors), amyotrophic lateral sclerosis (ALS) (up to 30% of patients have dysarthria as a first or predominant sign in the early stage of the disease, moreover a mixed speech impairment affects up to 70% of patients with limb onset at a later stage), Parkinson's disease (with nearly 90% of patients experiencing hypokinetic

dysarthria), multiple sclerosis (up to 50% of patients live with dysarthria) [16, 17, 18]. From these data, it is evident that dysarthria condition coexists with various motor disabilities, especially severe, that affect movement and coordination. In these conditions, the presence of dysarthria creates a complex interplay that impacts communication and overall quality of life for millions of people worldwide [19].

2.3 Acoustic features in disordered speech

People with atypical voices, particularly those living with dysarthria, exhibit considerable variability in their own articulatory output. This variability encompasses both intra-speaker variation (different speech produced by a single person) and inter-speaker variation (differences in speech across different speakers). A person with dysarthria may experience significant fluctuations in speech quality based on factors such as the time of day, stress levels, fatigue, and environmental conditions. Additionally, the severity of dysarthria and the involvement of various aspects of the speech production system contribute to substantial variability among speakers with non-standard speech.

In particular, three key mismatches exist between atypical and typical speech: irregular speaking rate, less distinct phone classes, and shifts in boundary positions. In most conditions, impaired speech typically exhibits a slower speaking rate. Damage to the neural-motor system in speakers with dysarthria makes it challenging to move articulators from one pronunciation position to another. Consequently, they require more time to produce utterances compared to typical speakers. Dysarthric speech also often exhibits less distinct vowel spaces. The authors of [20] found that the vowel space in dysarthric speech tends to be more centralized or overlapping, while typical speech shows greater distinctiveness. Reduced articulator flexibility in people with dysarthria contributes to this phenomenon. The issues in boundary position shifts refer to changes in the boundaries between voice and voiceless contrasts, in particular the standard category boundary positions in dysarthric speech shift to higher values compared to typical speech. These new boundaries are consistent for mildly dysarthric speech but inconsistent for more severe cases. Consequently, minimal pairs (e.g., voice and voiceless) maintain contrast in mildly dysarthric speech but may not always be distinguishable in more severe cases.In continuous speech, there are no clear breaks between individual words, making it the listener's task to detect word boundaries. Coarticulation, where neighboring sounds influence the acoustic realization of a phoneme, can create ambiguity at word boundaries. This ambiguity is exacerbated in dysarthric speech due to high phonemic uncertainty. Factors such as merging acoustic boundaries, indistinguishable voiced and voiceless contrasts, and different phone classes contribute to the slurred and

blurred quality of continuous dysarthric speech, further complicating the detection of word boundaries. Consequently, the automated recognition of continuous dysarthric utterances is more challenging than recognizing isolated single words.

To summarize, a speaker with dysarthria may exhibit a combination between the following characteristics:

- abnormalities in the rate of speech;
- changes in voice quality;
- irregular pitch and rhythm;
- inaccurate speech production leading to phoneme distortions;
- altered breathing mechanism;
- choppy speech that may be hard to comprehend, especially in non-mild severity conditions.

2.4 Challenges in disordered speech recognition

The aforementioned features affect speech intelligibility, creating communication barriers for individuals with disordered speech and, consequently, difficulties in social participation and independence. Due to dysarthria, atypical speech can be significantly unintelligible, causing a typical audience to experience issues communicating with dysarthric speakers unless they have prior experience with such individuals. For these reasons, speakers with dysarthria often face practical barriers in their daily lives, reducing their level of autonomy in everyday activities, such as making a phone call or participating in a videoconference. Additionally, many people with dysarthria experience severe motor disabilities, such as spastic quadriplegia, which hinder their ability to interact with computers and digital devices via traditional input interfaces like a mouse, keyboard, or touchscreen.

For such individuals, automatic speech recognition (ASR) technologies can be a desirable alternative to enable them to interface with digital devices or become a communication intermediary [3]. ASR involves the use of algorithms and models to process and interpret audio signals, identifying the words and sentences spoken by a user. Nowadays, these technologies can significantly improve the quality of life for individuals with these disabilities through their applications in the areas of communication and human-computer interaction. However, despite significant advancements in ASR technology, current systems remain inadequate for

speakers with disordered speech, especially in moderate to severe conditions. This inadequacy is due to multiple factors. The phonemes pronounced by individuals with atypical voices can be highly imprecise, with pitch pauses in vocalic segments and inaccuracies in consonant production. These alterations may also mask the discriminative acoustic attributes that ASR systems rely upon to recognize phonemes. Furthermore, because the effects of the disability vary from one individual to another, speech variations among dysarthric speakers are significantly greater than in normal speech. These differences make the acoustic modeling components in standard ASR systems ineffective in correctly mapping dysarthric speech signals to phonemes, as they need to address challenges caused by unusual and imprecise phonation, tempo, and speed inconsistencies.

Recent studies by Jaddoh et al. (e.g., [9, 21, 4]) have analyzed the challenges in natural speech interaction between users with atypical voices and speech recognition systems, particularly those powering voice assistants and conversational agents. The primary issue with ASR systems in this context is their poor performance, which tends to deteriorate for individuals with dysarthria, worsening as the severity of the condition increases. One of the main challenge is the accuracy of these systems in understanding commands issued with dysarthric speech. Variability in volume and pitch adds another layer of complexity. For instance, fluctuations in volume and pitch within a single word or sentence can confuse these systems, making it difficult to accurately capture the intended command. Additionally, the unique characteristics of dysarthric speech, such as breaths between syllables, further complicate the recognition process, as described in [9].

Different articles have utilized the TORGO corpus to explore the accuracy of dysarthric speech recognition achieved by three speech recognition cloud platforms [2]: IBM Watson Speech-to-Text, Google Cloud Speech, and Microsoft Azure Bing Speech. According to these studies, the ASR systems exhibit comparable performance in recognizing dysarthric speech, with accuracy closely tied to the speech intelligibility of the end user. Overall, the platforms struggle when dysarthric speech intelligibility is low (80-90% word error rate), but improve significantly, achieving a word error rate of 15-25% for individuals with typical speech. The authors of [6] collected speech samples from eight Italian users with ALS-induced dysarthria. Based on these audio contributions, the differences between the aforementioned virtual assistant services in terms of speech recognition and response consistency were examined. The final results show varying performances among the virtual assistants. For speech recognition, Google Assistant is the most promising, with a word error rate of around 25% per sentence. In terms of response consistency, Siri and Google Assistant provide coherent answers approximately 60% of the time. However, end users often need to reformulate their speech requests multiple times to receive a coherent response from the virtual assistant platform. This is

very frustrating for individuals with speech disorders and may prevent them from relying on such technology in daily activities. As a consequence, from a perspective of a user with the above disabilities, the popular speech-controlled interfaces represent new barriers rather than technological aids helping his / her daily life.

Furthermore, a key challenge in building accurate dysarthric ASRs lies in the inherent variability of speech. This variability necessitates a substantial amount of additional audio contributions to train ASR acoustic models effectively. However, acquiring sufficient dysarthric speech data is challenging due to the limited availability of such data from speakers with non-standard speech. Collecting audio contributions is a difficult and time-consuming activity, as people with dysarthria often struggle with vocalization tasks (especially repetitive activities) and live with physical impairments that make producing large amounts of speech tiring. Consequently, compared to the more widely available normal speech corpora, such as Mozilla Common Voice, which contains thousands of audio samples, existing disordered speech corpora are smaller in size. For example, UA-Speech, a widely used English dysarthric speech corpus, contains 102.7 hours of speech from 13 healthy control speakers and 16 individuals with dysarthria caused by cerebral palsy. Other examples of impaired speech corpora include TORGO (15 hours), Dutch EST (6 hours), Cantonese CUDYS (10 hours), and Nemours (2.5 hours). Moreover, annotating speech data from individuals with non-standard voice is particularly challenging, further increasing the difficulty of constructing specific databases. Most of these corpora contain multiple repetitions of isolated words because it is easier for speech-impaired people to articulate single words rather than continuous sequences of terms or connected speech. Similarly, this is more effective when the size of the ASR vocabulary is small and includes only simple words with one or two syllables to boost recognition rates by reducing or minimizing ASR error in the presence of atypical voices and dysarthria [22]. Therefore, isolated word and small vocabulary ASR models are in increased demand for impaired speech recognition.

2.5 Proposed solution

In this thesis, we intend to address and simplify the multiple challenges discussed previously. Within the framework of digital assistive technologies, we propose the design and implementation of a technological ecosystem for automatic speech recognition (ASR) tasks in the presence of speech impairments, with a special focus on dysarthria. By leveraging artificial intelligence techniques, particularly deep learning, our goal is to realize ASR services that recognize a closed set of Italian speech commands, including isolated words and short sentences, spoken by selected users with atypical speech who are Italian speakers. This

system is not designed for continuous speech recognition. We adopt a speaker-dependent approach, requiring end users to contribute personal voice samples to train our ASR platform. This effort aims to realize a Voice User Interface (VUI) that enables users with speech disabilities to interact vocally with digital devices and computers in a smart environment. Given that many people living with neuromotor disabilities typically use a limited number of discrete input commands (e.g., those assigned to a numerical keypad to control a computer mouse) to fully interact with digital devices, our main idea is to replicate this functionality through automated recognition of specific disordered utterances, without the need to interpret conversational speech. Following this research direction, we believe that, in the field of human-computer interaction, amplifying the residual speaking abilities of persons with disordered speech is essential to work toward a more inclusive AI designed to break down barriers for disabled people. Therefore, a synergy between assistive technologies and innovation in ICT plays a crucial role.

The proposed digital ecosystem is designed for people who are verbal and experience difficulties in interacting vocally with unfamiliar communication partners and with the today's speech-controlled interfaces available on AI-powered devices, such as popular voice assistants' services running on smart speakers and phones.

In particular, our infrastructure is centered on the Italian language and is built upon three key pillars. The first pillar addresses the issues in disordered speech collection: due to the absence of corpora containing impaired samples, especially in non-English languages, one of the crucial activities involves creating such a corpus in Italian. To support the collection of speech data from people with disabilities, our intervention builds on our previous work [23], where we introduced the use of our CapisciAMe software to facilitate atypical speech acquisition while end users are prompted to speak specific isolated words and sentences aloud. In conjunction with technological improvements to our application, this has allowed us to enrich our private database of Italian atypical speech in terms of both volume and speech data variability. At the same time, since the process of speech signal enhancement is of crucial importance in building performant speech recognition tools (especially those relying on limited amounts of labeled data), we have introduced a dedicated procedure to ensure reliable training data from disordered speech contributions. Such filtered data represent the inputs for the second pillar of the proposed ecosystem, which is concerned with the application of Transformer models to empower various speech recognition engines. This is a crucial aspect of our work because it allows us to empower the third pillar of our digital ecosystem, which deals with the development of assistive technology solutions based on our speech recognition services. In this scenario, the creation of the VUI is decisive in supporting innovative systems in the field of human-computer interaction for people with atypical voices,

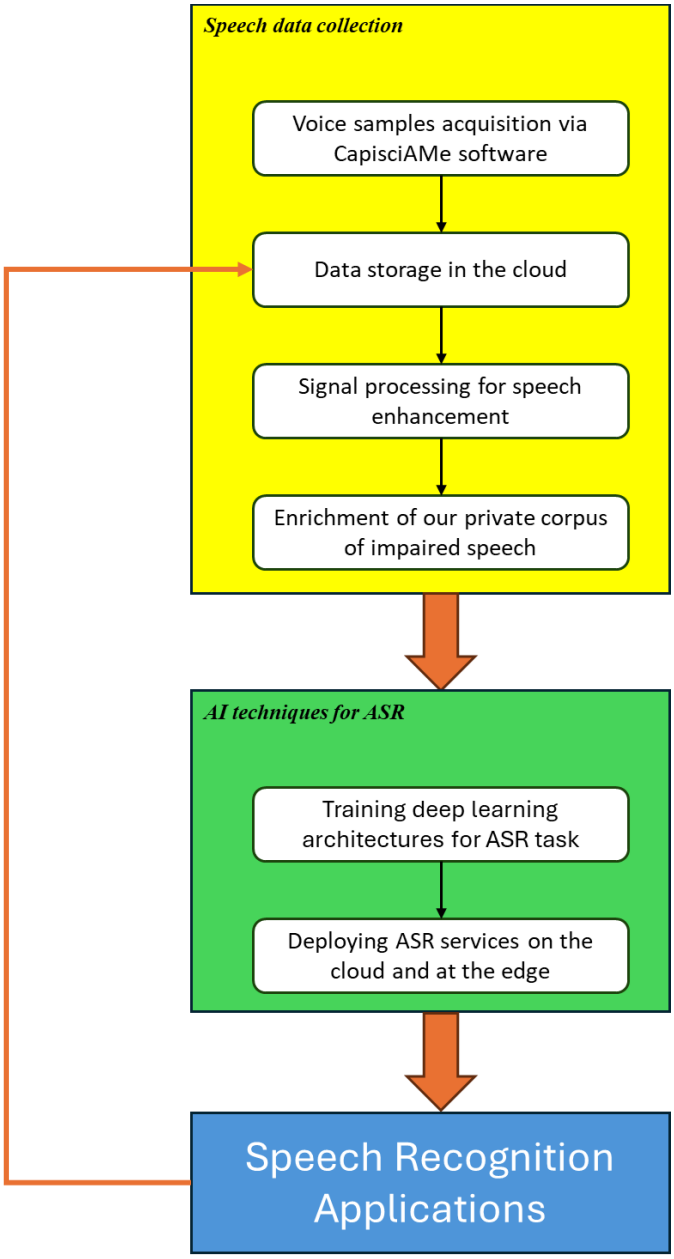


Figure 2.1: Key functional blocks of the proposed digital ecosystem for automatic disordered speech recognition

for example, in a customized smart home automation scenario. Additionally, the practical use of these real-world applications is crucial, as it allows us to initiate a virtuous cycle to encourage the collection of more voice data from people with speech disabilities who benefit from our ASR engine. It is also essential to deploy these software resources as a service over the cloud, to empower speech transcription services.

Figure 2.1 illustrates the main components of our technological ecosystem and the relationships between its key functional blocks. More specifically, the first block focuses on disordered speech data collection and outlines our strategy for acquiring voice samples from individuals who face challenges in interpersonal communication due to speech disorders. Unlike other research projects that collect speech data in collaboration with hospitals, clinics, and patient associations, our approach leverages social networks to directly reach people living with speech disorders. Our solution is also innovative from a technological perspective. It utilizes a common smartphone to acquire speech samples from end users through our CapisciAMe application. This free software allows users to easily record their utterances while speaking aloud a specific set of Italian isolated words and short sentences. The collected speech samples are then stored in our private cloud repository for subsequent processing. However, since personal mobile devices can be used in a variety of physical environments and situations (both supervised and unsupervised), many audio recordings may include dysarthric speech signals coupled with unwanted noises, such as bioacoustic, physiological, and background noise, resulting in poor overall quality. To address this, we manually review all collected voice recordings and conduct a dedicated speech signal enhancement process driven by a human listener. This step is crucial as it allows us to enrich our private corpus of Italian disordered speech (the CapisciAMe database, presented in Chapter 4) with verified and annotated speech recordings.

Thanks to this data availability, in the second block of our digital ecosystem, we apply AI-powered techniques for automatic speech recognition to train specialized deep learning architectures. In particular, while in our previous works we explored the utilization of convolutional neural network (CNN) structures for keyword spotting tasks in the presence of impaired speech patterns, the present thesis mainly investigates the impact of sequence-to-sequence models on disordered ASR. Our research exploits state-of-the-art ASR architectures that are already pre-trained on a significant quantity of typical speech (i.e., without dysarthria and other speech impairments) and then fine-tuned on our private database of disordered speech. Here, the application of a fine-tuning approach is critically important and enables us to move a key step toward the automated recognition of precise short sentences (as combinations of keywords) pronounced by selected users with speech disorders. At the same time, the ASR model inference process is crucial, as it, in conjunction with cloud-

based services, is a key enabler for the development of real-world applications that harness the power of speech recognition in multiple scenarios. In this context, we also propose solutions to execute the inference on edge computing nodes, which are designed to operate with local hardware-software resources, without the need to access the cloud to process the speech signals.

The third block of our technological ecosystem is concerned with the development of assistive technology applications based on our speech transcription services. We present some prototypes in multiple fields (such as human-computer interaction) designed to support daily activities. At the core of these solutions, the creation of a common Voice User Interface (VUI) to interact with digital devices is of paramount importance. At the same time, in order to mitigate the challenges in atypical speech recognition, it is essential that this system recognizes precise utterances spoken by users with voice disorders, without the need to work on continuous speech. As highlighted in Figure 2.1, the practical usage of the proposed applications plays a critical role, as it represents a fundamental way to empower the speech data collection from people who might benefit from our solution. Indeed, the collected speech samples are stored in our private cloud repository, ready for processing to enrich our corpus of disordered speech.

2.6 Thesis innovation

The present thesis aims to address the challenges in automatic disordered speech recognition from a technological perspective by proposing a synergy between deep learning-based techniques for ASR and digital assistive technologies. The main innovation of our work includes the following major aspects of the proposed digital ecosystem:

- Innovative strategies for collecting speech samples from speakers with dysarthria and atypical speech.
- Enrichment of our private corpus of Italian disordered speech, primarily designed to support AI-driven research.
- Application of state-of-the-art encoder-decoder models for automatic recognition of impaired speech in the Italian language.
- Creation of cloud-based ASR services for transcribing disordered speech.
- Development of voice user interfaces and applications that leverage the aforementioned ASR services in assistive technology scenarios

2.7 Summary

This chapter focused on the concept of speech impairment, with a special emphasis on dysarthria. We also discussed the major challenges in automatic disordered speech recognition and introduced the main components of our digital ecosystem that support ASR services for atypical speech.

Chapter 3

Literature review and state-of-the-art

The challenges of automated speech recognition in the presence of speech impairments, such as dysarthria, have garnered significant attention from both industry and research communities [24, 25, 26, 27, 28, 29].

In this chapter, we present relevant literature works in the field of disordered speech recognition and briefly discuss ongoing research projects based on collaborations between scientists and major companies. Before starting our analysis, the following section describes the features of major corpora (databases) containing speech samples from speakers with speech disorders.

3.1 Disordered speech corpora

Currently, the major corpora used in studies on automatic disordered speech recognition consist predominantly of speech collections in English or American. This creates an unwanted bias, as the potential benefits of ASR solutions based on such corpora may be limited to speakers of these languages. The following subsections present the characteristics of pre-existing impaired speech corpora.

3.1.1 The UA-Speech corpus

The Universal Access (UA)-Speech corpus was developed by researchers at the University of Illinois. It contains American speech samples from 16 speakers with cerebral palsy and dysarthria, and 13 age-matched participants with typical speech (i.e., without impairments). The speakers with dysarthria are categorized into four severity levels (severe, moderate-severe, moderate, and mild) based on a subjective estimate of perceptual speech intelligibility [30]. The UA-Speech corpus is widely used in literature, as it is the largest dysarthric speech

corpus in American English with well-defined training and test partitions. The collected data consist of isolated words. Specifically, each speaker produced a total of 455 single words, including 29 NATO alphabet letters, 19 command words (e.g., 'no', 'yes', 'up', 'down'), 100 common words, 10 digits, and 300 uncommon words from the "Grandfather" passage. The entire content is divided into three blocks for each speaker, with each block containing the common words and one-third of the uncommon words. Typically, in scientific publications, data from blocks 1 and 3 are used as the training dataset, while block 2 is used as the testing dataset. All these speech recordings were captured by an eight-microphone array sampled at 48kHz [31]. The UA-Speech database comprises of 102.7 hours of labeled speech and is not publicly accessible for general use. It is available upon request only to researchers at government and academic institutions.

3.1.2 The TORGO corpus

The TORGO corpus was developed by the University of Toronto's departments of Computer Science and Speech Language Pathology in collaboration with the Holland-Bloorview Kids Rehabilitation Hospital in Toronto. TORGO is a Canadian English corpus comprising 21 hours of synchronized acoustic and articulatory recordings from 15 speakers [32]. Among these, eight speakers (five males and three females) have varying degrees of dysarthria (ranging from severe to mild) due to conditions like cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). The remaining participants are typical speakers matched by age and gender (four males and three females). On average, each speaker with dysarthria recorded 415 utterances, while each typical speaker recorded 800 utterances. The audio recordings were captured using a head-mounted microphone and an array microphone, both sampled at 16 kHz. The recordings include four types of stimuli: non-words, isolated words, sentences, and photograph descriptions. Typically, only isolated words and sentences are used in experiments. Non-word recordings serve to establish baseline speaker abilities during the recording phase, and photograph descriptions lack text transcriptions. The isolated words set includes English digits, international radio alphabets, the twenty most frequent words in the British National Corpus, and some phonetically contrasting word pairs selected by [33]. Most of the restricted sentences were chosen from the Yorkston-Beukelman assessment of intelligibility [34] and the TIMIT database [35].

The TORGO corpus contains 615 unique words and 354 unique sentences, with a total vocabulary size of 1573 words. The vocabulary size for sentence prompts alone is 1083 words. Additionally, aligned articulatory data for some utterances were recorded using a 3-D AG500 electromagnetic midsagittal articulography system [36]. At the time of this writing, only a

portion of the TORGO corpus is publicly available [1].

3.1.3 The Euphonia corpus

Although the Euphonia corpus has not yet been made available for open access, its inclusion is warranted due to its unique contributions to addressing challenges found in other disordered speech datasets. Notably, the Euphonia corpus is the largest of its kind, comprising one million utterances and over 1300 hours of pathological speech. This extensive dataset was collected as part of Google's Project Euphonia, involving over 1000 participants who recorded their utterances [37]. Unlike other impaired speech datasets in English, Euphonia's recordings are highly realistic, as participants used a web application to record their speech in non-controlled environments (e.g., at home) with their own devices (e.g., cell phones and laptops), which introduced various technical and noise issues. While these factors can affect recognition accuracy during training, they help bridge the gap between training and generalization errors when deploying the trained ASR model [38]. The substantial size of the Euphonia dataset also allows for a comprehensive representation of various speech impairments, including those associated with conditions like cerebral palsy, Down syndrome, and stroke.

3.1.4 The homeService corpus

The homeService corpus is a British English collection of dysarthric speech data recorded in home environments (approximately 10 hours of total recording time). It aims to develop a system that functions effectively in real-world settings to assist individuals with dysarthria in interacting with electronic devices using single-word commands [39]. The database contains speech from five speakers (three males and two females) with severe dysarthria. Among them, three have cerebral palsy (CP), one has motor neurone disease (MND), and the condition of the fifth speaker is unspecified. The data collection involved two methods: enrolment data (ER) and interaction data (ID). ER data was gathered as participants read lists of chosen command words, while ID data was recorded as participants used the homeService speech-enabled interface to operate household devices. Three speakers provided both ER and ID data, whereas the other two only contributed ER data due to personal reasons. The total recording duration is approximately 10 hours, with a vocabulary size of 131 words. The audio recordings were sampled at 48 kHz.

[1] https://catalog.ldc.upenn.edu/LDC2012S02

3.1.5 The Nemours corpus

The Nemours corpus is a collection of English dysarthric speech data featuring 11 male speakers with cerebral palsy [40], each exhibiting varying degrees of dysarthria as assessed by the Frenchay Dysarthria Assessment (FDA) tool. Each speaker contributed 74 short nonsense sentences and two paragraphs of connected speech, specifically from the Grandfather [41] and Rainbow passages. In total, the corpus includes 814 short nonsense sentences. The audio recordings are sampled at 16 kHz.

3.1.6 The Whitaker corpus

The Whitaker English corpus comprises 19275 isolated-word utterances from six individuals with varying degrees of dysarthria caused by cerebral palsy [42]. Additionally, it includes utterances from a typical speaker for reference purposes. The vocabulary is divided into two sets: the TI-46 set, which includes 46 words (the alphabet, digits, and 10 control words), and the Grandfather set, which consists of 35 phonetically diverse words.

3.1.7 Other corpora

In the related literature, there are many examples of impaired speech databases in non-English languages. Yan et al. [43] present the Chinese Dysarthria Speech Database, the most extensive collection of Chinese disordered utterances to date, featuring 133 hours of recordings from 44 speakers with various levels of speech disabilities. The Netherlandic Dutch dysarthric corpus [44] encompasses approximately six hours of speech data from patients suffering from Parkinson's disease, traumatic brain injuries, and cerebrovascular accidents. Various pre-designed word and sentence sets have been used to ensure diverse and balanced phonetic and lexical content. French atypical speech data has been collected in the DesPho-APaDy project [45], resulting in two corpora that include recordings of 860 and 699 dysarthric speakers, respectively. Furthermore, the COPAS corpus of pathological speech in Flemish Dutch [46] contains recordings from 75 dysarthric speakers with conditions such as multiple sclerosis and Parkinson's disease. This collection was realized within the scope of the SPACE project, aiming to develop a reliable ASR-based speech assessment tool for pathological speech. Other examples of atypical speech corpora include the Czech CLARIN collection [47], the TYPALOC corpus in French [48], the Korean impaied speech collection [49] and the Cantonese dysarthric corpus [50].

Apart from our CapisciAMe corpus, recent literature identifies two disordered speech datasets in Italian: IDEA [51] and EasyCall [52]. The IDEA corpus includes a total of

16,794 speech recordings from 45 speakers affected by eight different pathologies. Since this corpus is not publicly available, we cannot evaluate our methodology on it. The EasyCall corpus is a public database of command speech in Italian, recorded from healthy individuals and dysarthric patients. Its organization does not match our methodology because it considers isolated words and short sentences, resulting in a limited number of impaired speech recordings per class. Additionally, our preliminary analysis of the EasyCall corpus has revealed significant transcription errors that negatively impact the approaches proposed in our research project.

3.2 Related works

Scientific contributions about interactions between people with disordered speech, particularly dysarthria, and speech technologies have been recognized since the early 1980s, e.g., [53]. Several papers have investigated how speech intelligibility contributes to ASR performance over the years. Ferrier et al. [54] discussed the importance of speech recognition as a method for speakers with spastic dysarthria to create digital documents, and they studied the correlation between speaker intelligibility and recognition success, finding a strong correlation between intelligibility and ASR performance [22]. In recent years, with the advent of cloud computing, these systems are deployed on devices like smart speakers and phones, utilizing cloud-processed data for enhanced performance. However, these systems often fail to work on moderate to severe impaired speech due to imprecise consonant articulation, addition or deletion of phones, reduced loudness, hypernasality, or other characteristics associated with speech impediments. Although normal speech ASR can be used to recognize mild dysarthria [55], various researchers have investigated ASR systems specifically designed to recognize dysarthric speech, as detailed in the following subsections.

We present a literature analysis in the field of automatic disordered speech recognition, discussing various techniques. These include traditional solutions based on Hidden Markov Models (HMMs), deep learning-inspired solutions using artificial neural networks, sequence-to-sequence (Seq2Seq) models based on transformer, and state-of-the-art ASR architectures [56, 57] pre-trained with large datasets of typical speech and fine-tuned on smaller corpora of non-standard speech.

3.2.1 Traditional ASR solutions

These ASR systems primarily rely on Hidden Markov Models (HMMs), which are statistical models representing systems with hidden states. HMMs are particularly effective in speech

recognition due to their ability to model time series data and represent phonemes. One notable attempt to design dysarthric-specific ASR using conventional methods is investigated in [58]. The authors proposed a vector HMM approach for a dysarthric ASR system with a vocabulary of ten digits and 196 common words. This system was evaluated by a 44-year-old male with mixed spastic athetoid cerebral palsy, achieving an intelligibility score of 59.3%. Additionally, two individuals with CP had intelligibility scores of 65% and 22%, as determined by the Computerized Assessment of Intelligibility of Dysarthric Speech (CAIDS). The reported recognition rates ranged from 78% to 95%.

Another example is the study by Rajeswari and Chandrakala [59], which proposed a Generative Model-Driven Feature Learning approach for dysarthric speech recognition. This method aimed to recognize isolated words from a partition of the UA-Speech database, which includes a 29-word vocabulary. However, HMM-based ASR systems have become less popular due to several limitations. HMMs assume that all probabilities depend solely on the current state, ignoring correlations between subsequent input frames. This assumption poses challenges for processing sequential acoustic frames in speech tasks. Additionally, HMM probability density models often have suboptimal accuracy, and the Maximum Likelihood training requirement can lead to poor discrimination among acoustic models, especially when training data for dysarthric speech is limited [60].

As a result, the scientific community has shifted towards deep learning-based approaches, which can overcome many of the limitations associated with HMMs.

3.2.2 ASR solutions based on Artificial Neural Networks

Nowadays, ASR approaches utilizing Artificial Neural Networks (ANN) have become more popular due to their flexibility and ability to learn complex speech patterns, leading to high-performing modern ASR systems.

Among the earlier solutions for impaired speech recognition based entirely on ANN is the study in [61]. Shahamiri et al. investigated the design of a speaker-dependent tool based on the Multi-View Enhanced Multi-Learner active learning theory [62], which utilized an array of neural networks, each tasked with learning the acoustic features of words in the vocabulary. In a different work, nine UA-Speech subjects were randomly selected, and their speech was used to train a Recurrent Neural Network (RNN) dysarthric ASR with a 100-word vocabulary [63]. Acoustic features were presented via Glottal to Noise Excitation (GNE), indicating whether a given voice signal originates from vibrations of the vocal folds or turbulent noise generated in the vocal tract. The mean accuracy reported in this study was approximately 85%, although the authors did not provide per-subject performance, which is

common in dysarthric ASR research due to the high variability of dysarthric speech.

Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for dysarthric ASR were also studied, for example in [64]. Verifying on the Nemours corpus, the study reported that their CNN achieved up to 82% accuracy for speakers having mild speech impediments. Hybrid models that use different algorithms for various ASR components have also been studied to improve dysarthric ASR performance. In this context, the author of [65] proposed a hybrid model combining Hidden Markov Models (HMM) and neural networks, evaluated by three cerebral palsy subjects. The study reported a mean accuracy of 97% for each subject over a small vocabulary. Additionally, remarkable results on the UA-Speech database were delivered by Speech Vision [3] a speaker-adaptive dysarthric ASR based on Spatial Convolutional Neural Networks. This system was verified by all UA-Speech subjects over a 155-word vocabulary. Unlike other dysarthric ASR systems, Speech Vision learns to recognize the shape of words pronounced by dysarthric subjects using a voicegram presentation of speech and leverages transfer learning and synthetic generation of dysarthric speech to overcome the data scarcity problem.

The effectiveness of CNN models in disordered ASR has recently been confirmed for the Italian language as well. In [66], a word recognition accuracy of approximately 85% was achieved using a CNN model trained with only a subset of our Italian speech collection. Similar results were found in [67, 68, 69], where the evaluation involved a small number of participants with CP-induced dysarthria, using a different version of our private corpus of atypical speech, which is also used in the present PhD thesis.

3.2.3 Sequence-to-sequence ASR systems

Sequence-to-sequence (Seq2Seq) models were developed to address problems involving sequences of unknown lengths [70]. Initially used for machine translation, such techniques are now applied to various sequence modeling tasks. In a Seq2Seq model, one recurrent neural network (RNN), i.e., the encoder, processes the input to create a vector representation, which another RNN, i.e., the decoder, uses to generate the output. Seq2Seq models have become popular in the speech community for their ability to convert input sequences to output sequences. Deep learning frameworks are well-suited for this task due to their large capacity and end-to-end training capabilities, allowing direct mapping from input signals to target sequences [71, 72]. Initially, Seq2Seq ASR systems were commonly built using RNNs. For instance, the authors of [73] proposed a speech ASR where both the encoder and decoder components were RNNs. Another example is [74], which employed a deep bidirectional RNN to encode the speech signal into a suitable feature representation and an

attention-based Recurrent Sequence Generator RNN to decode this representation into a sequence of characters. However, using RNNs as the primary algorithm imposes limitations that prevent Seq2Seq ASR systems from reaching their full potential [75], While RNNs work well for short statements, their ability to learn larger contexts is limited due to their shorter reference window. Additionally, RNNs' sequential nature makes them very slow to train. To overcome these limitations, Transformers and self-attention were proposed by Vaswani et al. [76] in 2017. The attention network intuitively learns to focus on important features and ignore the rest, making the features context-aware. With self-attention, the network can generate representations for characters based on surrounding characters, modulating token representations. Attention acts as an interface between the encoder and decoder, providing the decoder with information from the encoder's hidden states. With multi-head attention, the attention operation can be performed multiple times for each attention layer. Transformers are deep neural networks that leverage the self-attention concept, commonly used in modern Seq2Seq tasks. A study by Karita et al. [77] compared transformers and RNNs and concluded that transformers learn faster and deliver better inference results. Another contribution [78] in the text-to-speech context observed that a larger minibatch resulted in better validation L1 loss for transformers with faster training but had a negative influence on the L1 loss for RNNs. Transformer and attention-based ASR were first introduced in [79], where a 2-D attention mechanism was proposed and evaluated on the Wall Street Journal speech corpus. This study resulted in significantly lower training costs and achieved an excellent Word Error Rate (WER), demonstrating the Speech Transformer's efficiency and efficacy. Since their introduction, transformers have been utilized in various ASR systems. Regarding Seq2Seq dysarthric ASR, Google's Project Euphonia researchers employed a RNN-T architecture [80] composed of an encoder with eight LSTM layers and a language model with two LSTM layers [81]. This ASR was evaluated on the Euphonia corpus [38], which includes 294 dysarthria participants among other speech-impaired subjects. The same authors indicated that impaired speech was particularly difficult to model and hence was classified among high word-error-rate subsets [82]. Furthermore, a Transformer-based application in automatic speech recognition can be found in [30]. The authors experimented with proposed two Seq2Seq architectures supporting an ASR tested on the UA-Speech corpus. In particular, the researchers experimented and measured how increasing the depth and number of transformer encoders and decoders could lead to better performances, and how using depthwise separable convolution instead of fully-connected encoder components could improve word recognition accuracies. from healthy speakers (without a speech disorder).

3.2.4 Fine-tuning approaches in disordered speech recognition

In the realm of deep learning, fine-tuning is a technique where a pre-trained model is adapted to perform a new task. Instead of training a model from scratch, which can be both computationally expensive and time-consuming, fine-tuning leverages the knowledge acquired from a large dataset on a related task. This technique involves adjusting the model parameters to fit the hypothesis space of the target task using a much smaller amount of data compared to pretraining. It is typically employed when the source and target domains are similar.

Numerous studies have explored the use of fine-tuning techniques to enhance model performance. Well-known self-supervised speech models such as the Wav2vec series [57], Hubert [83], WavLM, and MMS (Massively Multilingual Speech) [84] require fine-tuning on domain-specific data to adapt to downstream tasks. The authors of [85] explored various pretraining and fine-tuning methods for the Wav2vec 2.0 model on the ASR task for child speech. Similar contributions can be found in [86, 87, 88, 89] Zhang et al. analyzed different combinations of pretraining and fine-tuning on 15 low-resource languages in the OpenASR21 challenge. Pasad et al. utilized various metrics [90], including canonical correlation analysis, mutual information, word recognition, and word similarity, to study and analyze the characteristics of Wav2vec 2.0's layer representations and guide model improvements for better fine-tuning strategies.

Recent works have highlighted the advantages of Wav2vec2 and fine-tuning approaches in automatic speech recognition (ASR) for atypical speech [91]. A PhD dissertation investigated the application of this model, along with the XLSR variant, for recognizing Dutch dysarthric speech [92]. However, it did not show significant improvements in word recognition accuracy compared to a custom supervised model based on Deep Neural Networks and Hidden Markov Models (DNN-HMM). Baskar et al. emphasized the necessity of speaker-dependent auxiliary features, such as fMLLR and xvectors, to adapt Wav2Vec2 models for better dysarthric speech recognition [93]. They also conducted cross-lingual experiments using English and German datasets [91]. A contrastive learning SSL approach [94] was explored with data augmentation to enhance impaired speech representations. The researchers used transfer learning, pre-training the model on non-dysarthric speech, and found that contrastive learning SSL outperformed supervised pretraining. These results were primarily observed in American and Japanese languages [95].

Hasegawa-Johnson et al. investigated the impact of Wav2Vec2 fine-tuning in the presence of Parkinson's disease (PD)-induced dysarthria [10], using a subset of the Speech Accessibility Project (SAP) database. The authors demonstrated that a Wav2Vec2-base system fine-tuned with speakers with PD (the SAP training set) could recognize a different set of texts, spoken

by a different set of speakers with PD (the unshared test set associated with the same dysarthric speech corpus), with an error rate of 23.69%. This error rate is less than two-thirds of the error rate achieved on the same data by a system fine-tuned using 960 hours of speech from people without PD.

Other research contributions, such as [96], suggest training an acoustic model with features extracted from cross-lingual models like XLSR, Hubert, and Wav2Vec. These studies demonstrate that pre-training speech representations on large unlabeled datasets can improve word error rate (WER) performance. Positive outcomes were noted in English speakers with CP-induced dysarthria (UA-Speech corpus) and Spanish speakers with Parkinsonian dysarthria (PC-GITA corpus). Additionally, applications in aphasic speech recognition are discussed in [97], while Sanguedolce et al. [98] propose using OpenAI's Whisper in a rehabilitation scenario involving patients with post-stroke aphasia. Other authors employed Whisper's architecture for automatic disordered speech recognition in English: for instance, Rathod et al. highlighted a WRA of 59% using a block of 155 keywords belonging to the English UA-Speech corpus [99]. Vinotha et al. [100] propose fine-tuning the Whisper model for impaired ASR by including additional features extracted from Mel-frequency cepstral coefficients (MFCCs). By combining spectrograms and MFCCs within an attention mechanism, the model creates a richer feature representation, with spectrograms providing broader context and MFCCs highlighting crucial formant frequencies. This methodology has been tested on UA-Speech corpus with an average word recognition accuracy of 74.08%.

As shown in Table 3.1, the application of fine-tuning methods in the automated recognition of Italian disordered speech remains unexplored nowadays. To fill this gap, one of the major contribution of this thesis is to investigate the utilization of including state-of-the-art sequence-to-sequence ASR models, exploiting encoder-decoder techniques, for the automatic recognition of atypical speech in Italian.

3.3 Current ongoing projects

Nowadays, there is increasing public interest in tackling the global challenges of automatic disordered speech recognition. Collaborative efforts among diverse researchers and stakeholders are essential in addressing these issues effectively. For instance, by considering the American language, the Speech Accessibility Project is a collaborative initiative led by the University of Illinois Urbana-Champaign, in partnership with major tech companies like Amazon, Apple, Google, Meta, and Microsoft. This project aims to enhance voice recognition technology to better serve individuals with diverse speech patterns, including those affected by neurological conditions such as Parkinson's disease, cerebral palsy, and Down

Table 3.1: Articles on the utilization of fine-tuning approaches in impaired speech recognition

Study in	**Language**	**Method**
[100]	English (UaSpeech corpus)	Whisper model with MFCC features extraction
[99]	English (UaSpeech corpus)	Whisper and Bi-LSTM classifier model
[10]	English (SAP corpus)	Wav2Vec2 base model
[98]	English (SONIVA corpus)	Whisper large model
[93]	English (UaSpeech) German (private corpus)	Fne-tuning Wav2Vec2 using fMLLR features
[95]	English (UaSpeech) Japanese (ELSpeech corpus)	Wav2Vec2 + Wav2LM
[101]	English (UaSpeech and DementiaBank corpus)	SSL pre-trained Wav2Vec2 with hybrid TDNN and Conformer
[91]	Dutch (private database)	Various Wav2Vec2 and Whisper variants
[97]	English and Spanish (AphasiaBank database)	Wav2Vec2 XLSR-53 model

syndrome [2]. By collecting and utilizing a vast array of speech samples from these communities, the project seeks to reduce the error rate in speech recognition systems, making them more inclusive and effective. This initiative not only addresses the current limitations of voice recognition technology but also promotes greater accessibility and independence for people with dysarthria and other speech disorders. The current phase of the project focuses on individuals living with Parkinson's disease [10], with plans to expand the dataset in the future to include other types of speech impairments.

Google Euphonia and Google Relate are closely connected initiatives aimed at improving communication for individuals with non-standard speech. Google Euphonia is a research project that focuses on enhancing speech recognition technology by collecting and analyzing voice samples from people with speech impairments. This project aims to train models that can better understand and transcribe atypical speech patterns, making voice-activated technologies more accessible. The data and insights gained from Project Euphonia are crucial in developing more accurate and inclusive speech recognition systems, which directly benefit applications like Google Relate for Android devices. In particular, this software is designed to assist users in real-time communication. By leveraging the advancements made through Google Euphonia, Google Relate can more effectively understand and transcribe the unique speech patterns of its users. The app offers features [102] such as restating spoken words in a

[2] https://speechaccessibilityproject.beckman.illinois.edu/

clear, synthesized voice, live transcription, and integration with Google Assistant for various tasks. Essentially, Google Relate applies the research and technology developed by Google Euphonia to provide practical, everyday solutions for individuals with speech impairments, enhancing their ability to interact with technology and communicate more effectively.

3.4 Summary

This chapter provided a literature review in the field of automatic disordered speech recognition, discussing various techniques such as traditional solutions based on Hidden Markov Models, deep learning-inspired approaches using artificial neural networks, sequence-to-sequence models based on transformers, and state-of-the-art ASR architectures that support fine-tuning. We also highlighted the main features of dysarthric corpora available in the literature.

Chapter 4

Disordered speech data collection

This chapter deals with the first pillar of our digital ecosystem for automatic speech recognition in the presence of dysarthria and other speech disorders. It tackles a significant challenge in creating speech recognition solutions for atypical speech: the limited availability of datasets featuring speech samples from individuals with speech disabilities, which are crucial for training speech models, especially those based on deep learning strategies.

Nowadays, in the presence of pathological speech, the issue of audio data scarcity is exacerbated. Collecting impaired audio contributions is a difficult and time-consuming task, as individuals with speech impairments often struggle with vocalization tasks (especially repetitive activities) and may have physical impairments that make producing large amounts of speech tiring. Consequently, compared to the more widely available normal speech corpora, such as Mozilla Common Voice, which contains thousands of audio samples, existing disordered speech corpora are significantly smaller. This leads to a bias in the datasets used to train today's ASR solutions that power popular virtual assistants and conversational agents. As a result, this underrepresentation contributes to excluding individuals with speech disabilities from benefiting from current voice recognition technologies, which are not robust to their atypical speech patterns. From the perspective of users with these disabilities, speech-controlled interfaces represent new artificial barriers rather than technological aids supporting their daily activities.

To address these issues, we believe that introducing assistive technologies and novel methodologies aimed at facilitating speech acquisition from individuals who are verbal and experience disordered speech is of paramount importance. For these reasons, our research proposes the usage of mobile devices, such as smartphones or tablets, to collect atypical voices. We leverage the widespread use of smartphones among users with disabilities to distribute our CapisciAMe software, available for free. This software is designed to encourage individuals with speech disorders to share, or donate, their personal utterances in the form

of speech recordings while they speak predetermined voice commands, including a limited set of single isolated words and short sentences. Through the app, all recorded data can be submitted to our private cloud repository, contributing to the enrichment of our private database used in training speech models for automatic impaired speech recognition [23].

The rest of this chapter describes the main features of our CapisciAMe software and provides details on our private corpus of disordered speech in Italian.

4.1 The CapisciAMe solution

The digital ecosystem discussed in this thesis is named CapisciAMe, an Italian dialect expression meaning "Understand Me". This name reflects an invocation from Davide, the author of this dissertation, who has lived experience with cerebral palsy and severe dysarthria, to popular virtual assistants that fail to recognize his speech commands. The CapisciAMe mobile application is a key component of the ecosystem, primarily designed to facilitate the collection of speech samples from individuals with speech disorders who struggle to interact with voice assistants and to communicate with unfamiliar communication partners.

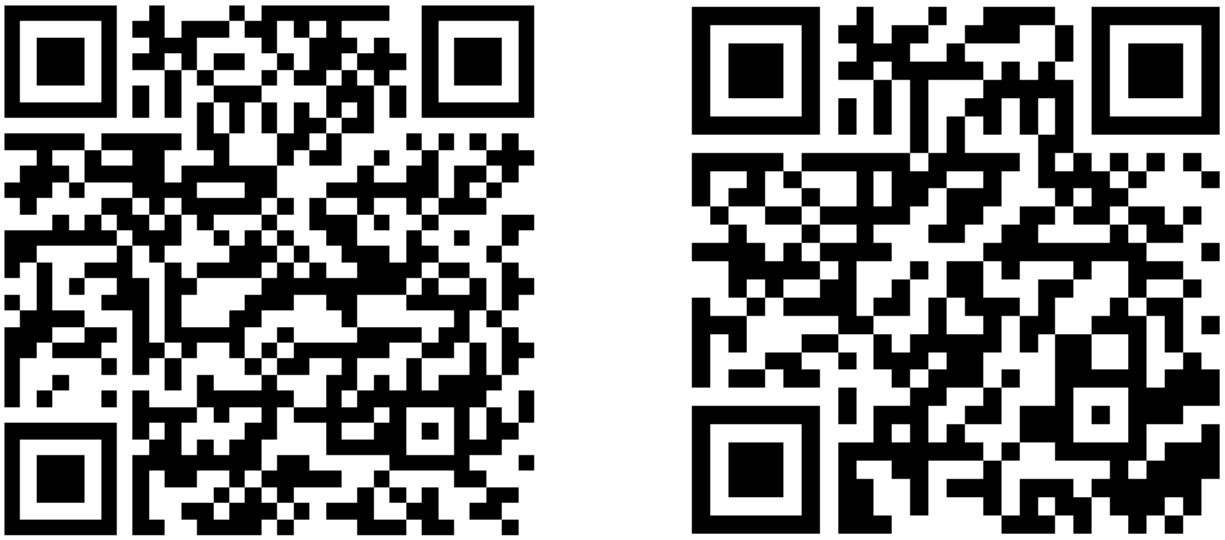
Unlike other studies that suggest using a personal computer to collect dysarthric speech in a supervised setting [29], our research focuses on using a common smartphone. This device integrates all necessary hardware and software components for recording voice samples and can interface with external microphones to achieve high-quality recordings. The portability of the smartphone allows users to collect personal utterances anytime and anywhere. Our solution can be used both autonomously by the speech-impaired person and in a controlled setting, such as a rehabilitation context, where a specialist can monitor the speech acquisition process [68].

However, some users with severe motor disabilities may find a mobile device challenging to use, particularly the touch screen. For these users, we are developing a desktop application that replicates the main functionalities of CapisciAMe on a personal computer.

In the following, we present the key features of our CapisciAMe application for mobiles.

4.1.1 The mobile application

CapisciAMe is a free native application available for both Android and iOS devices (see more details in Figure 4.1), currently localized in Italian and English. The app enables anonymous interaction with our digital ecosystem. Users with speech disabilities log in using a personal nickname, without needing to provide additional personal information such as the cause or severity of their speech disorders, although they have the option to specify these details



(a) Get the app on Google Play Store

(b) Get the app on AppStore

Figure 4.1: QR codes to get the CapisciAMe app

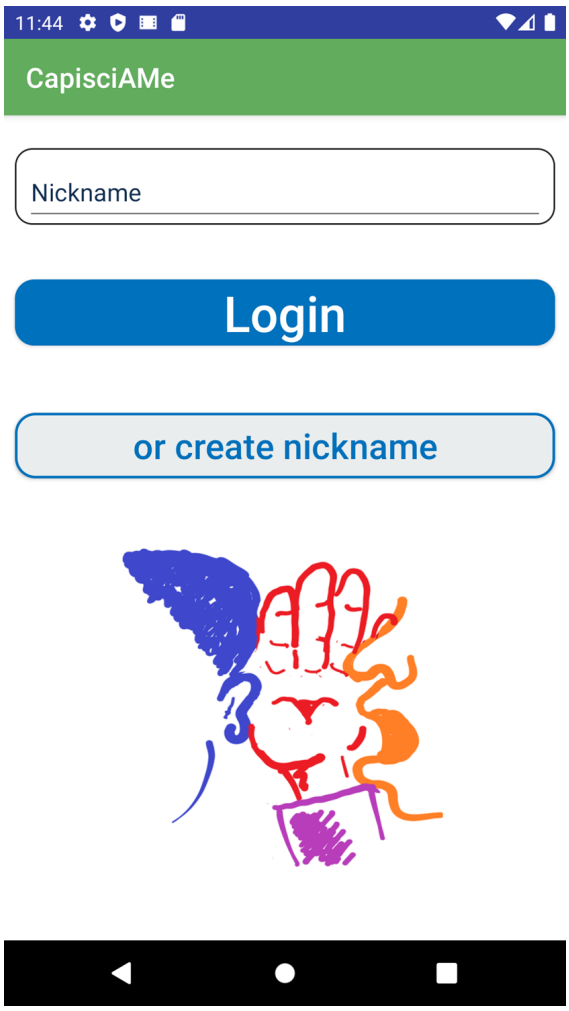
if they wish, as reported in the registration form (see Figure 4.2b). This approach makes privacy optional for our users.

As depicted in Figure 4.4, our app consists of two main sections: Training and Recognition. The Training section guides users through the repetition of relevant isolated words and sentences to enhance our private dataset of impaired speech. The Recognition section offers a convenient way to test the speech recognition capabilities of our ASR engine. This is done using a speaker-dependent approach, where users can verify the ability to recognize the speech commands used to train the ASR system. Both sections of the app provide four different modalities to suggest the input voice command to the user, as shown in Table 4.1.

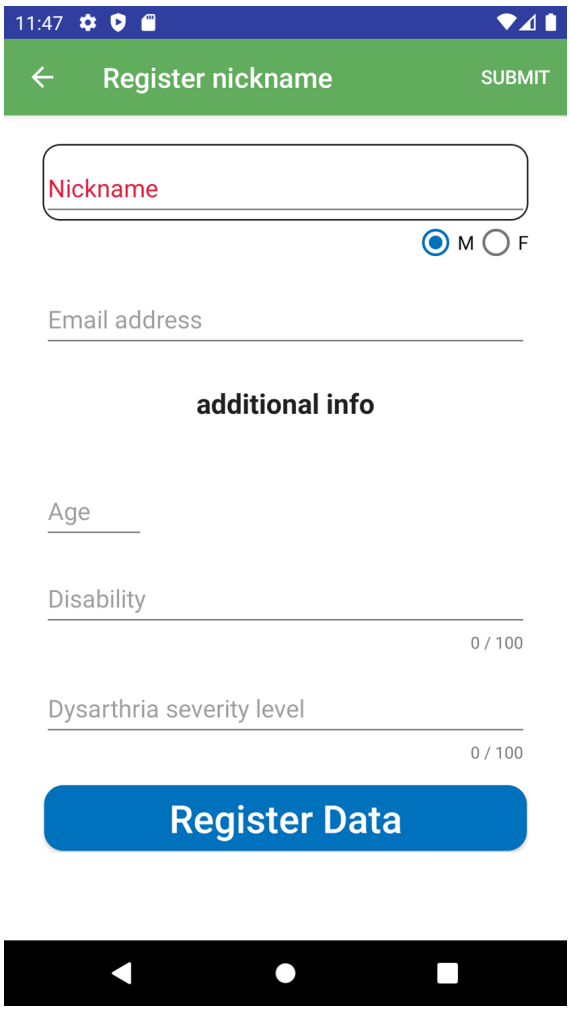
These modalities include combinations of the following elements:

- Text caption: The speech command is displayed on the smartphone's screen within a label, as shown in Figure 4.5a.
- Text-to-speech: The speech command is read aloud by the app via a text-to-speech service.
- Pictures: The speech command is suggested through the visualization of a picture, as described in Figure 4.5b.

The application of the aforementioned methods is crucial as it helps us obtain annotated impaired speech recordings, i.e., audio contributions. The CapisciAMe application allows users to select which terms they wish to train or attempt to recognize. Examples of such



(a) Login screen

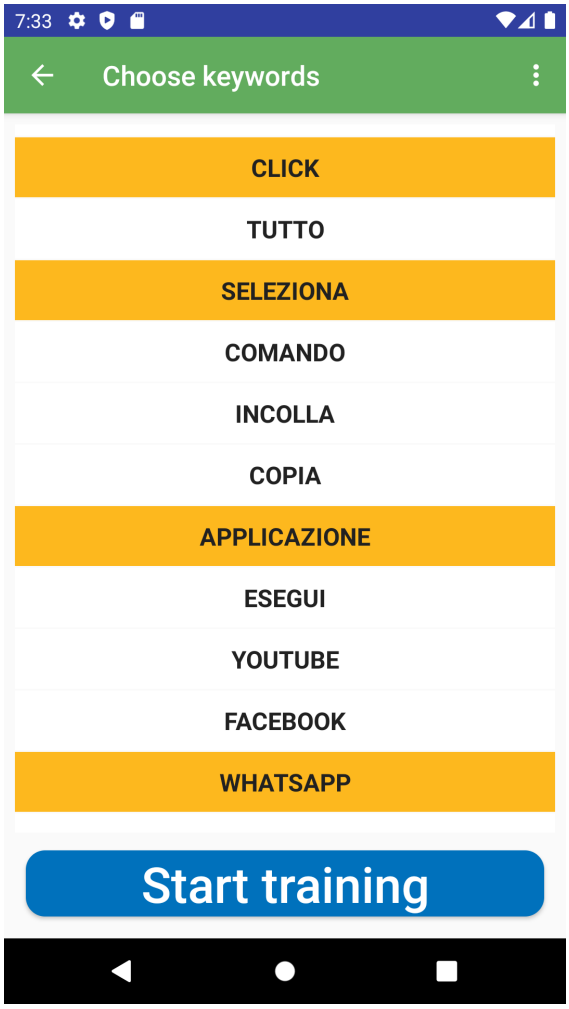


(b) Registration form

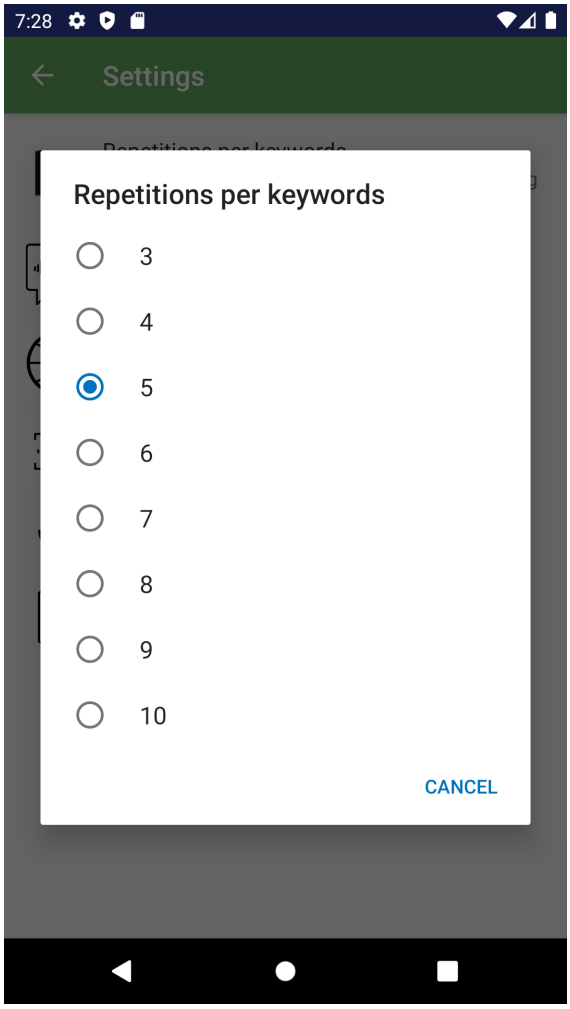
Figure 4.2: CapisciAMe app: login screen and registration form

	Elements		
Modalities	**Text caption**	**Text-to-speech**	**Pictures**
1	**X**	**X**	
2		**X**	
3			**X**
4	**X**		

Table 4.1: Modalities to suggest speech commands in CapisciAMe



(a) Examples of keywords list



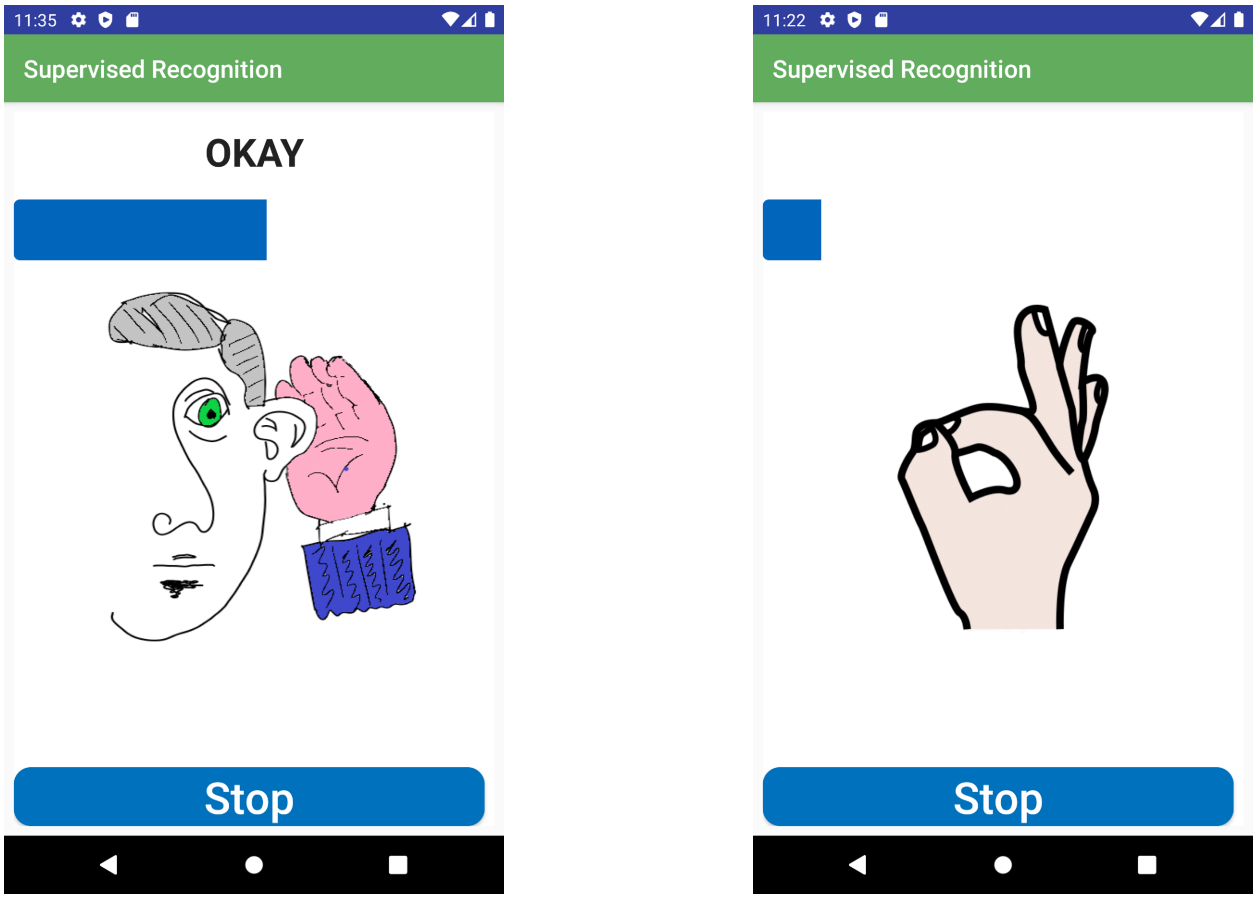
(b) Settings

Figure 4.3: CapisciAMe app: keywords list and settings

Figure 4.4: CapisciAMe app: home screen

features and settings are provided in Figure 4.3. Furthermore, the app integrates a standard speech recognition feature, where the software interprets the user's speech and converts it into a computer-generated voice. In this sense, CapisciAMe can function as a Voice Input Voice Output Communication Aid (VIVOCA), an augmentative and alternative communication (AAC) device that recognizes the disordered speech of the user and translates it into synthesized speech. To support these speech recognition applications, capturing the user's speech as they begin to speak is of paramount importance. Therefore, the CapisciAMe software integrates a basic Voice Activity Detection (VAD) algorithm to start and stop the acquisition of the speech signal from the microphone. Moreover, suitable audio feedback is provided to the user to guide the vocalization process, particularly during training sessions.

At the time of this writing, the main features available on the CapisciAMe mobile application are also included in a desktop application for Windows personal computers. This software is currently under development and is primarily designed for users who have difficulty using a smartphone, such as those with motor or visual impairments. As summarized in Table 5, one of the key advantages of using CapisciAMe in a desktop environment is that it can benefit from a wide range of assistive technology solutions available on traditional personal computers.



(a) With a text-caption

(b) With a picture

Figure 4.5: CapisciAMe app: two modalities to suggest the "okay" keyword to the user

	CapisciAMe on mobiles	**CapisciAMe on personal computers**
PROS	• The extensive adoption of smartphones and mobile devices by individuals with disabilities. • Capability to collect personal speech samples at any time and place. • Smartphones are equipped with all necessary hardware and software for voice recording. • Ability to connect a smartphone to an external, high-quality microphone for improved audio recording. • Speech recording can be a manageable activity for individuals who are weak in vocalisation tasks. • Users can utilize a wheelchair smartphone holder for easier interaction with the app.	• In a supervised environment, such as a rehabilitation center, a specialist can monitor the quality of audio recordings. • Higher quality speech samples can be achieved by using high-quality microphones paired with desktop or laptop computers. • CapisciAMe users may benefit from a wide range of assistive technology solutions available for desktop computers. • A computer can serve as an aid for speech therapy, allowing the acquisition of speech samples to be integrated into therapy sessions [68]. • In speech therapy contexts, many patients can be actively engaged in collecting voice samples.
CONS	• The app may be used in noisy environments. • Users might record speech samples incorrectly, leading to systematic errors that affect audio quality.	• Computer software may be more complex for end users compared to mobile apps. • Using a computer requires a fixed setting or specific environment, and individuals with disabilities may need supervision. • Computer software may not effectively motivate individuals with disabilities to donate their speech samples. • Software installation issues may necessitate specialized support.

Table 4.2: Benefits and drawbacks of CapisciAMe versions

4.2 Speech signal enhancement

The objective of speech enhancement techniques is to reduce background noise in speech signals, thereby improving the robustness and accuracy of automatic speech recognition (ASR) systems. Speech recordings made in real-world environments, such as homes, often contain unwanted components like ambient noise, music, traffic, wind, or other people's conversations. These factors can significantly degrade the quality of the speech signal, making it difficult for ASR systems to accurately recognize utterances. Consequently, noise suppression is of paramount importance and can generally be achieved through various methods, such as spectral subtraction, Wiener filtering, or more advanced techniques like deep learning-based models [103].

Within the framework of the CapisciAMe digital ecosystem, specialized noise reduction plays a fundamental role in obtaining reliable impaired speech input data for empowering subsequent ASR model training. At the time of this writing, the majority of the data (over 80 percent) in our corpus have been collected via our mobile app and recorded using the smartphone's integrated microphone, primarily within household environments. Consequently, the impaired speech information content in these recordings is accompanied by other audio signals that are not of interest and may negatively impact the entire ecosystem. Such unwanted components can include background noise, household noise, bioacoustics, and physiological signals, in conjunction with the inherent dysarthric speech features.

In most cases, voice recordings come from CapisciAMe users who have weak speech production abilities (especially in repetitive vocalization tasks). These recordings are characterized by poor articulation due to irregular sub-glottal pressure, loudness bursts, phoneme elongation, stereotyped vocalizations, and unexpected pauses during utterances, which may also not contain the expected speech command. Additionally, reduced motor skills and other debilitating conditions, including the user's psychophysical state, may randomly affect the input data. Other potential disturbances can arise from the caregiver's actions as a facilitator for dysarthric speech repetition, as the presence of a personal helper is crucial for supporting the voice signal acquisition process in many disability conditions.

Therefore, the use of denoising and speech enhancement techniques is mandatory to attain adequate performance in automatic disordered speech recognition, as motivated in the following research article:

- Mulfari, D., Campobello, G., Gugliandolo, G., Celesti, A., Villari, M., & Donato, N. (2022, June). Comparison of Noise Reduction Techniques for Dysarthric Speech Recognition. In 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA) (pp. 1-6). IEEE. DOI: 10.1109/MeMeA54994.2022.9856486

In more detail, the published research utilized a reduced portion of our private corpus of atypical speech in Italian to investigate the performance of a manual semi-empirical speech enhancement method compared to two automatic denoising tools. These tools are based on a classical audio denoising filter and a more recent deep learning approach for noise reduction using RNNs. Our analysis was based on three different strategies including:

- Classical Audio Denoising Filters. Several tools provide audio denoising filters based on (Fast Fourier Transform) FFT. In particular, the FFmpeg suite, an open-source and widely used software designed for multimedia stream management, includes several built-in noise suppression filters. Among these filters, the AFFTDN was selected for testing on our voice dataset. AFFTDN is a conventional noise suppressor based on the FFT algorithm, offering several parameters to set the noise floor, select the noise reduction in dB units, enable noise tracking, among others.

- RNN-based Denoising Filters. Among the available deep recurrent neural network models aimed at noise suppression, we considered RNNoise, a hybrid approach [104] that combines classic signal processing with deep learning techniques. It consists of an RNN with a total of four hidden layers, combined with a classical pitch filter that attenuates noise between pitch harmonics. Technically, the RNNoise algorithm has been integrated into the FFmpeg software.

- Manual Approach. We compared the above approaches with a manual semi-empirical speech enhancement procedure conducted by a human listener familiar with issues concerning speech intelligibility in the presence of dysarthria and other speech impairments. The human listener uses their sensitivity and experience to detect and remove unwanted acoustic components, such as audio glitches or caregiver's speech, from the atypical voice contributions. In particular, dysarthric speech within the audio track has been isolated from unwanted contributions and a simple gain equalization has been performed every time that the recorded speech signal level was too low. Furthermore, the human-driven intervention is essential to guarantee effective impaired speech annotation, which is of paramount importance for the training of deep models.

Figure 4.6 shows the waveform and the corresponding spectrogram of a speech signal representing the Italian keyword "sinistra" recorded using the smartphone's integrated microphone with the CapisciAMe app. Figures 4.7, 4.8, and 4.9 illustrate the application of the AFFTDN, RNNoise algorithm, and manual approach, respectively, on the same disordered speech signal. These speech enhancement techniques are not exclusively designed for impaired speech; their impact on the performance of an early version of the CapisciAMe ASR

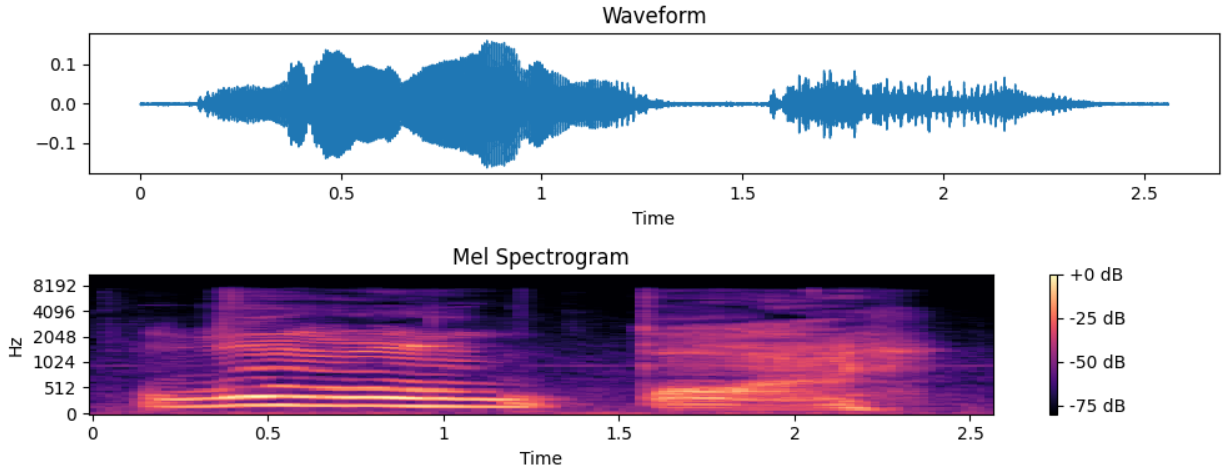


Figure 4.6: Example of a speech signal recorded with our app, without any filters applied

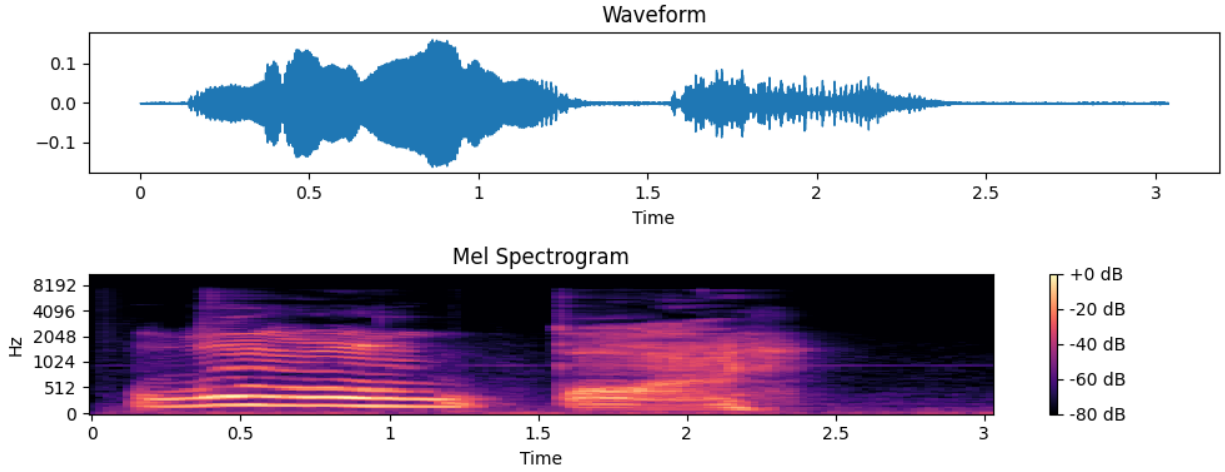


Figure 4.7: Example of a speech signal recorded with our app and filtered using AFFTDN

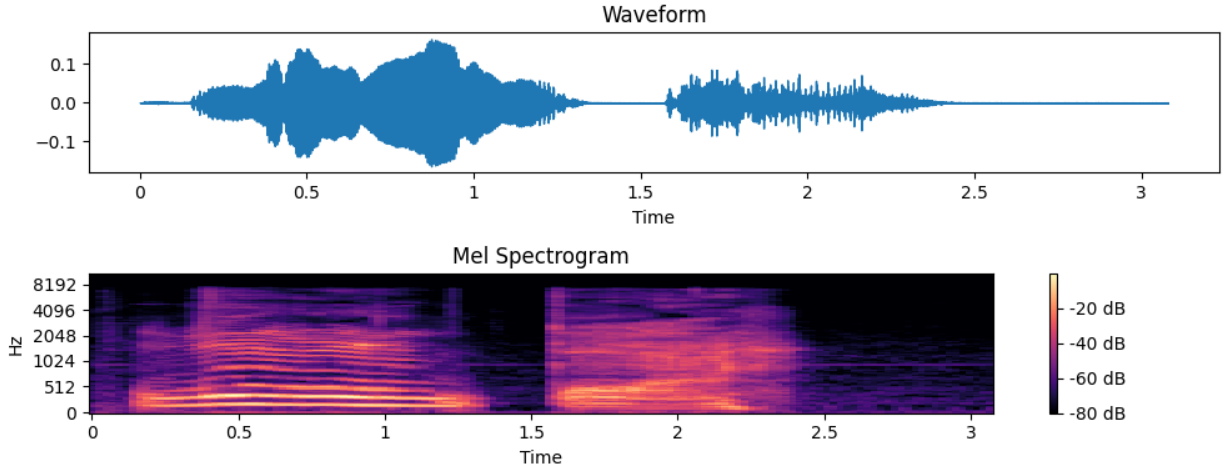


Figure 4.8: Example of a speech signal recorded with our app and filtered using the RNNoise algorithm

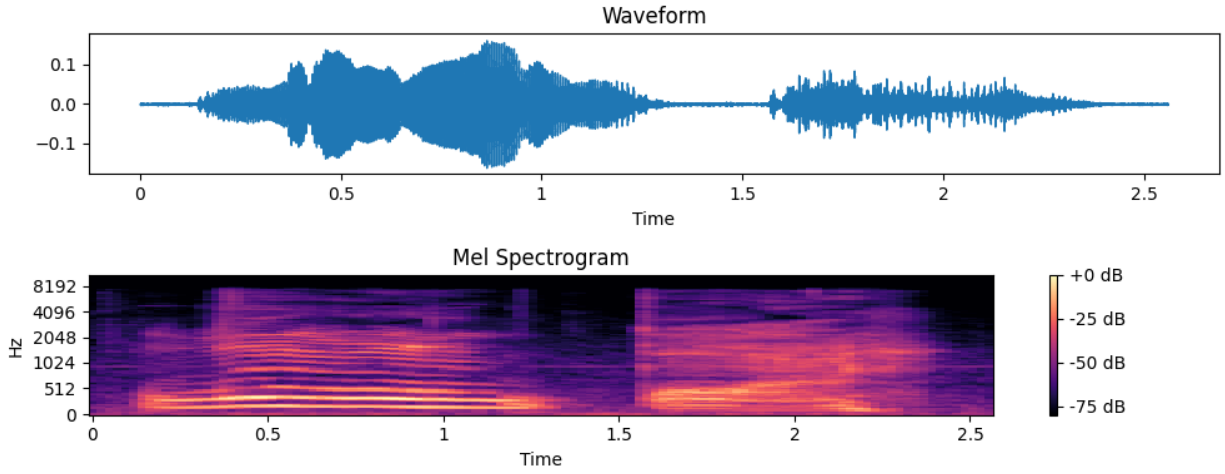


Figure 4.9: Example of a speech signal recorded with our app and filtered using our manual approach

system was evaluated in [105]. Specifically, this study focused on isolated word recognition within an ASR dictionary composed of thirteen Italian keywords, utilizing a deep learning system with a two-layer convolutional neural network for keyword spotting tasks. In the described conditions, a subset of our actual Italian speech database was used, divided into two separate partitions with no overlapping elements: the training set and the test set. The test set included 778 voice recordings split into 13 classes. These audio data were collected from seven CapisciAMe users (4 male and 3 female adults) with severe (3 persons) and moderate (4 persons) levels of dysarthria. Among these participants, six have infant cerebral palsy, and one adult is affected by cerebropathy. The training set included the remaining 21,184 speech audio contributions from 156 CapisciAMe users and was employed to train the cnn-trad-fpool3 architecture. Each audio contribution contains a single example of speech commands uttered by a selected speaker with dysarthria, and it has been stored in a separate uncompressed WAV audio with the following properties:

- Duration: 2560 ms
- Sample rate: 16 kHz
- Bit rate: 256 kps
- Number of channels: 1 (mono)
- Sample size: 16 bit

.

The testing dataset content was processed with three previously used denoising approaches, resulting in three corresponding denoised datasets. Batch inference processes were then executed to calculate the accuracy, i.e., the ratio of correctly predicted words to the total number of words, obtained on the three distinct testing datasets. Quantitative results expressed in terms of accuracy percentages are shown in Figure 4.10, where the "None" condition refers to the application of no filters or speech enhancement processes on the impaired speech data. More precisely, the accuracy achieved with the spectral denoising method AFFTDN is lower even compared to the "None" case (78.1% vs 79.2%), while the use of RNN provides only a modest improvement, while a significant improvement was achieved with limited manual audio preprocessing. Using the proposed semi-empirical manual approach, the achieved accuracy level reaches up to 97.0%.

We interpret this result as indicating that denoising tools erroneously considered dysarthric utterances as background noise, removing actual information content. This is because the

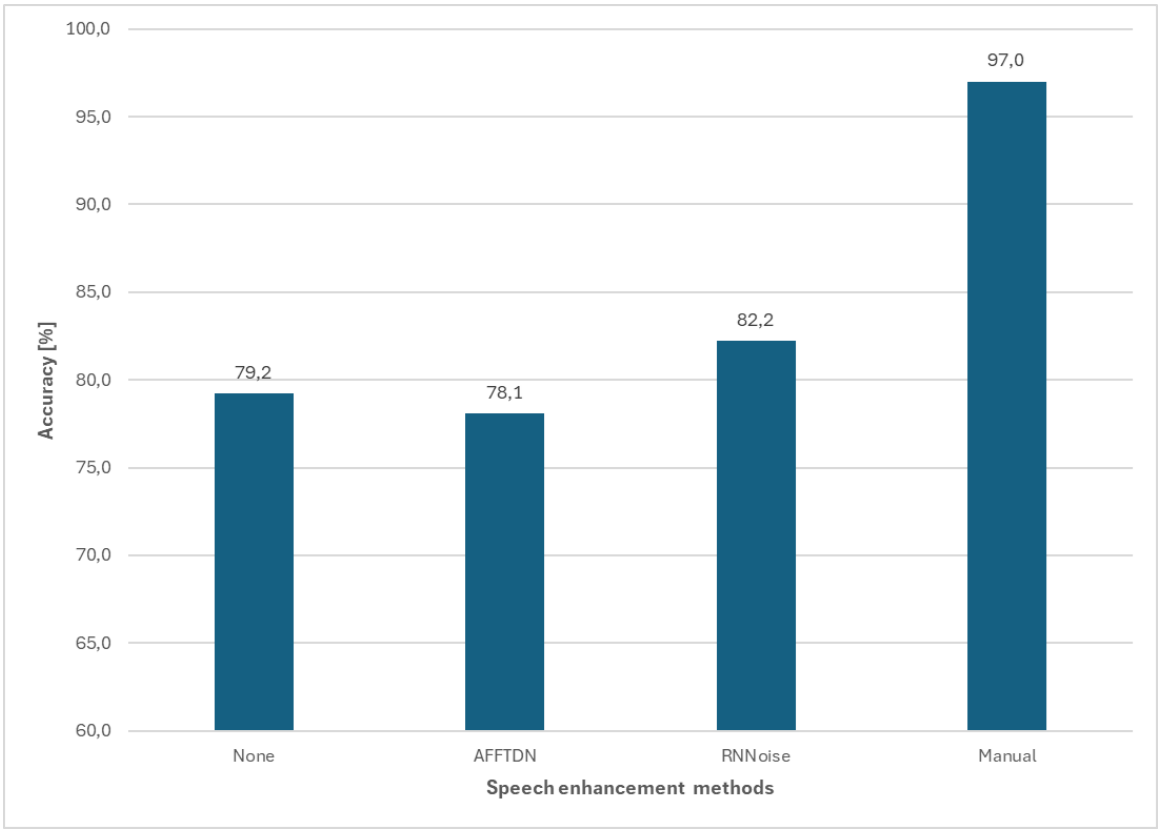


Figure 4.10: Comparison of denoising and speech enhancement techniques for automatic disordered speech recognition

disordered speech is often accompanied by poor articulation of consonants rather than vowels. Therefore, unvoiced consonants sound very similar to background noise. By listening to the audio files, we discovered several other reasons that justify the failures of automatic denoising tools, such as: i) stereotyped vocalizations in user's speech; ii) unwanted vocalization due to the end user's fatigue in speaking; iii) multiple examples of the same speech command in the input recordings; unvoiced segments at the beginning of the track.

As a result, contemporary standard speech enhancement and denoising techniques, which typically perform well with speakers with a standard speech, often irreparably corrupt the recordings authored by speech-impaired speakers. In contrast, manual intervention correctly suppresses, filters, or leaves these utterances unaltered, achieving an accuracy of up to 97.0%.

While the manual approach is crucial for ensuring denoised and reliable disordered speech contributions for ASR model training, its primary drawback is its inapplicability in real-time speech recognition scenarios. Consequently, we have implemented a dedicated software pipeline for speech signal enhancement, which encompasses the following steps:

1. Voice Activity Detection: Since the speech recognition system must operate continuously, it is critical to capture the user's speech as soon as they begin speaking. Therefore, a background process has been implemented to analyze the real-time input buffer from the microphone and compute its amplitude relative to the background noise.

2. Noise Reduction: It is important to separate human speech from unwanted signals, such as household noise, bioacoustics, and physiological signals. Initially, a band-pass filter of 200Hz - 4000Hz is applied to the atypical voice signal. Subsequently, using the "noisereduce" Python library, a more efficient noise reduction is conducted by leveraging a "spectral gating" technique to reduce the background noise and help isolate the disordered speech information.

3. Speech Signal Alignment: At this step, the audio signal from the previous step is converted to the WAVE format, with appropriate properties to integrate with our digital ecosystem.

Figure 4.11 illustrates the effect of the aforementioned steps on a single recording of a sentence pronounced by a person with dysarthria, under the form of an acoustic waveform and spectrogram. In the original waveform (see Figure 4.11a), the spectrogram reveals scattered noise components and less distinct speech patterns, highlighting the presence of background noise and the challenges in recognizing dysarthric speech. The impact of our pipeline is shown in Figure 4.11b, where the spectrogram displays reduced noise levels and clearer, more continuous speech patterns. This indicates that the pipeline has effectively

enhanced the speech signal, making it easier for ASR systems to recognize. Our speech enhancement algorithm successfully preserves and isolates the atypical speech information while suppressing other components that are not relevant for speech recognition purposes.

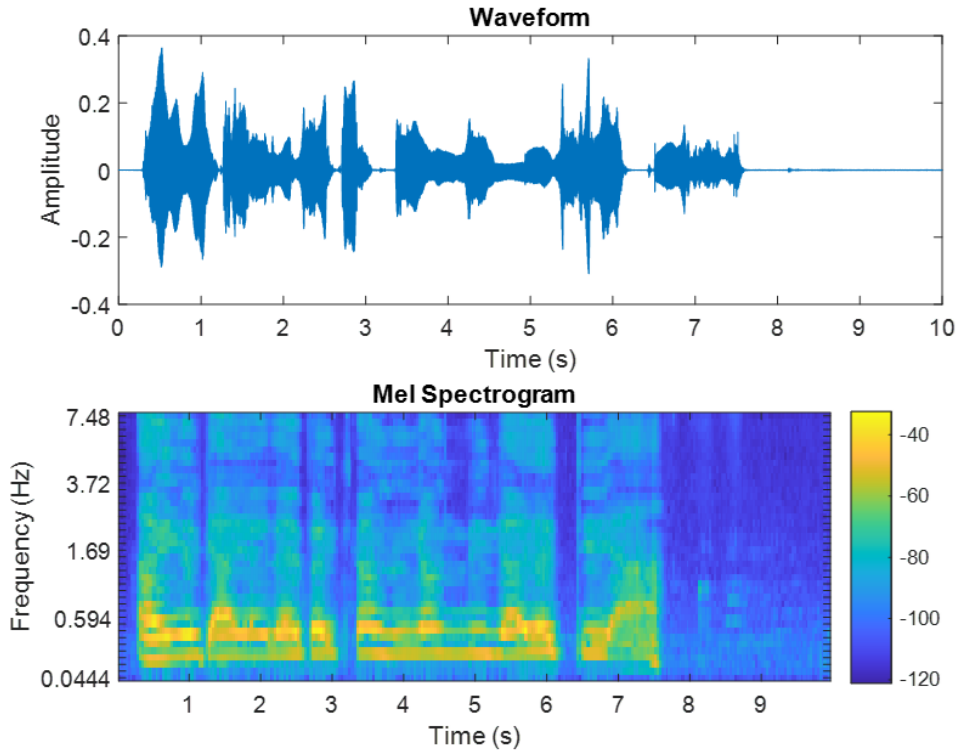
Furthermore, the effectiveness of the above software pipeline has been evaluated in comparison to the RNNoise algorithm previously discussed. Specifically, under the same conditions as the cited research, applying our software pipeline to the same testing datasets resulted in a 5.7% improvement in performance, achieving an overall accuracy of 87.9%. Consequently, the software pipeline has been integrated into the transcription services for disordered speech, which will be detailed in Chapter 6.

4.3 The CapisciAMe database

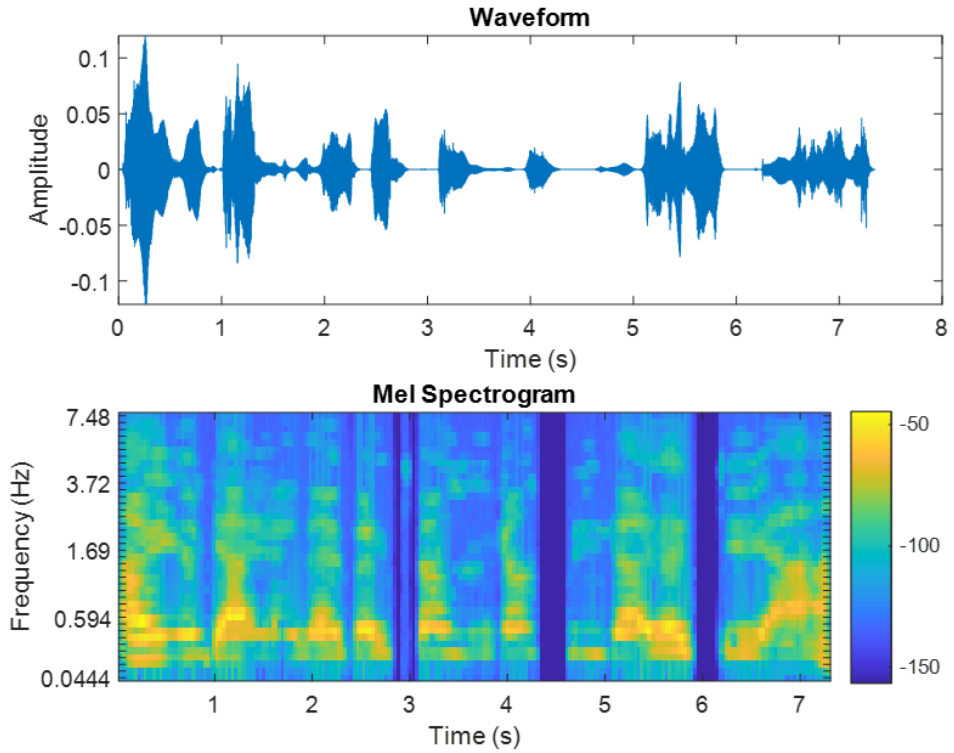
The construction and enhancement of our private corpus of impaired speech, i.e., the CapisciAMe database, are crucial for the research presented in this dissertation. This work aims to address a significant gap in current AI technologies, where speech patterns produced by individuals with dysarthria and other speech disorders are underrepresented. This lack of representation leads to a bias in today's speech datasets used for training AI models that serve current speech recognition systems. This issue is language-independent and represents one of the main reasons for the inadequate performance of speech-controlled interfaces on atypical voices. Therefore, addressing this bias [106] is essential, making data collection a crucial component of our digital ecosystem.

Currently, the acquisition of impaired speech through our CapisciAMe software is an ongoing activity conducted with the collaborative effort of individuals who experience speech impairments and small associations of persons living with neuromotor disabilities. We have leveraged the power of social networks and media to reach many users with speech disorders who wish to contribute spontaneously to our project. These users act as voice donors and are Italian speakers who face challenges in interpersonal communication due to their atypical speech.

As detailed in Figure 4.12, over the past few years, we have observed a remarkable increase in the total size of our speech corpus. Concurrently, we have enriched our limited ASR dictionary, i.e., the number of single keywords managed by our speech recognition system. At the beginning of our research, the CapisciAMe dictionary included just 13 terms [23, 105, 107, 68]. Gradually, thanks to improvements in convolutional neural network model architectures, our system was able to recognize up to 54 isolated Italian words [67, 66, 108]. More recently, with the introduction of state-of-the-art sequence-to-sequence models, the size of our ASR vocabulary has further increased. This was achieved through the adoption



(a) Original speech signal



(b) Enhanced speech signal

Figure 4.11: Impact of our pipeline for speech enhancement on a dysarthric voice recording.

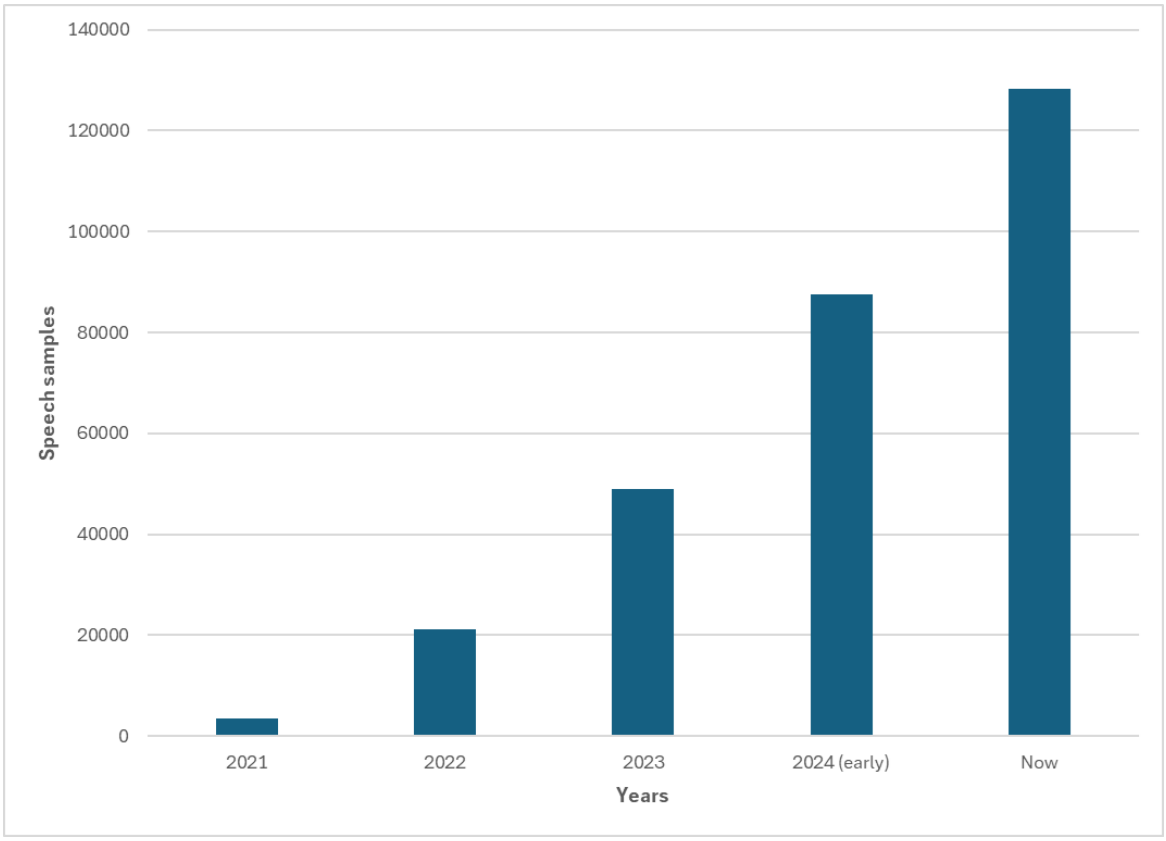


Figure 4.12: CapisciAMe database size across years

of fine-tuning approaches, where encoder-decoder architectures, such as OpenAI's Whisper and Meta's Wav2Vec2, are fine-tuned on our entire database of atypical speech in Italian [109, 110]. This represents a crucial improvement in our research activity in impaired ASR, leading to the opportunity to work on short sentences composed of a small number of Italian keywords, i.e., isolated words.

4.3.1 Structure

The CapisciAMe database is our private corpus of disordered speech in Italian, and its content is not publicly available. The structure of the corpus reflects the progression of our

Property	Value
Sample rate [kHz]	16
Number of channels	1
Sample size [bit]	16
Bit rate [kbps]	256

Table 4.3: Properties of the speech recordings in CapisciAMe database

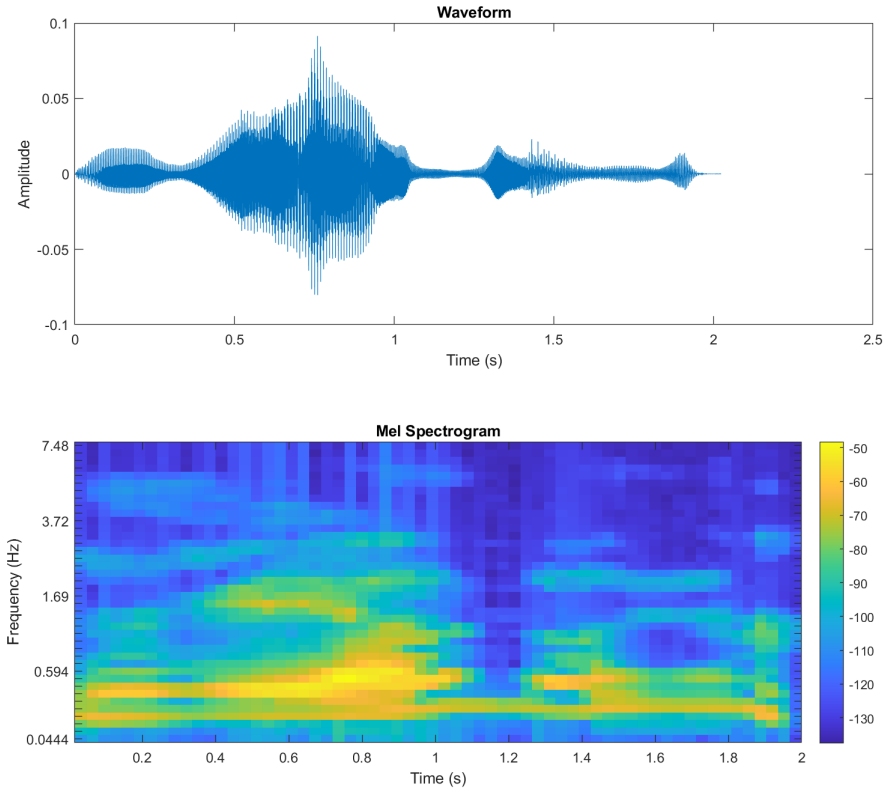


Figure 4.13: Example of a isolated word (keyword: "microfono") in our speech database

research activities over the past few years. As of this writing, we have collected two distinct annotated datasets:

- The first dataset contains a dictionary of 103 isolated words that encompass common terms used to express basic personal needs, such as requests for help, numbers from zero to ten, various colors, and commands to control smart home devices (e.g., plugs, lamps, televisions), play music, and interact vocally with smartphones and computers. Figure 4.13 shows an example of an isolated word, i.e., keyword, in our corpus in the form of an acoustic waveform and spectrogram.

- The second dataset contains a total of 4,632 distinct short, meaningful sentences. Each sentence is a combination of isolated words from the first dataset. Figure 4.14 provides an example of a sentence in our corpus in the form of an acoustic waveform and spectrogram.

As a result, the CapisciAMe database is the union of the above speech datasets and now encompasses more than 90 hours of disordered speech, corresponding to 128,246 individual

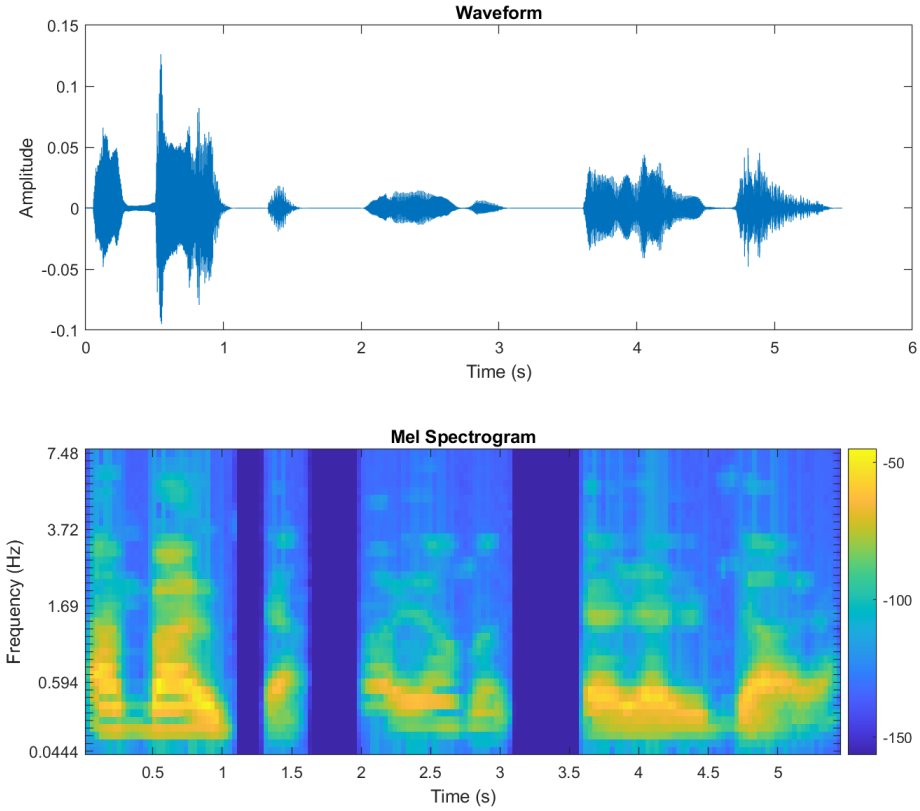


Figure 4.14: Example of a short sentence ("abbassa volume televisore") in our speech database

Feature	**Details**
Total Duration	91 hours and 11 minutes
Total Recordings	128,246 samples
Isolated Words	103 distinct elements
Word Recordings	106,522 samples
Short Sentences	4,632 distinct elements
Sentence Recordings	21,724 samples
User Population	250 anonymized users
Gender Distribution	151 male adults, 99 female adults
Language	Italian
Speech Types	Isolated words and short sentences

Table 4.4: Summary of the CapisciAMe database features

speech recordings. Each recording features a single pronunciation of a word or a short, meaningful sentence produced by a person with atypical speech who speaks Italian. Specifically, we have collected 106,522 speech recordings for words and 21,724 for short sentences. Each recording has been manually revised and filtered to ensure the robustness of the entire corpus. Our population includes a total of 250 anonymized users, comprising 151 male adults and 99 female adults. It is worth noting that our speech database is fully annotated and includes the transcription of each speech sample. Table 4.4 provides a summary of the key features of our disordered speech dataset in Italian.

Another key aspect of our work is that we do not rely on data augmentation, a technique commonly used in machine learning tasks that applies random oscillations and perturbations to augment the training data without affecting the class labels. In this regard, we do not generate artificial dysarthric data, setting our approach apart from recent studies that generate synthetic disordered speech signals, e.g., [111].

These unique features currently make the CapisciAMe database the richest and most complete corpus containing voice samples of people with speech disorders in Italian, to the best of our knowledge.

4.4 Summary

This chapter detailed our methodology for disordered speech collection, which represents a key pillar of our research. We described the innovative features of our CapisciAMe software. Over the years, this has led to the creation and enhancement of our private corpus of disordered speech in Italian, whose features are provided in Section 4.3.

Chapter 5

Deep learning approaches in disordered speech recognition

This chapter addresses the second pillar of our digital ecosystem, focusing on the utilization of various deep learning (DL) strategies to construct our speech recognition engine.

Deep learning is a subset of machine learning that employs artificial neural networks with multiple layers, known as deep neural networks, to simulate the complex decision-making processes of the human brain [112]. Unlike traditional machine learning models that rely on manual feature extraction, deep learning models automatically learn to extract features from raw data through a hierarchy of layers. This capability allows them to handle large amounts of unstructured data, such as images, audio, and text, making them particularly powerful for tasks like image recognition, natural language processing, and automatic speech recognition.

Within the framework of disordered ASR, we employ a deep learning-driven end-to-end recognition approach, where a single model directly maps the acoustic features of an audio signal to its corresponding textual representation. This method holds promise for understanding impaired speech patterns, offering advantages such as reduced computational demands and potentially improved performance [100]. Deep learning models, especially Convolutional Neural Networks (CNNs) and Transformer-based architectures [113], have demonstrated superior performance in recognizing and interpreting speech patterns compared to traditional methods. They can capture the intricate features of disordered speech, leading to higher accuracy in recognition tasks, as highlighted in recent studies [3, 30]. Another significant benefit is automatic feature extraction. Deep learning models can automatically extract relevant features from raw audio data, which is particularly advantageous for disordered speech, where manual feature extraction can be challenging due to the variability and complexity of speech impairments. Also, deep learning models are robust to the variability

in speech patterns caused by different types of speech disorders. They can learn to generalize from diverse speech samples, making them more adaptable to new and unseen data. This adaptability is crucial for developing robust ASR systems that can handle a wide range of speech impairments. Scalability is another advantage of deep learning. These models can be trained on large datasets, which is essential for developing robust ASR systems. As more speech data becomes available, these models can continue to improve, enhancing their performance over time [57].

However, as discussed in the previous chapters, one of the major challenges in developing speech recognition for pathological speech is the lack of available data, which hinders the generalization of ASR performance among speakers, especially those living with moderate and severe speech disorders. Consequently, ASR development for atypical voices can be reframed as the task of building a model with a limited amount of data. So, our CapisciAMe project leverages deep learning solutions and methodologies that can work with limited training datasets of labeled impaired speech. Specifically, two classes of deep learning models for ASR have been explored:

- Convolutional neural network models for isolated word recognition tasks. These are designed to predict the presence of a reduced number of speech commands (single keywords) within an atypical speech signal.

- Sequence-to-sequence structures that exploit sophisticated pre-existing DL frameworks, known as encoder-decoder models, which currently are state-of-the-art (SOTA) architectures, namely Wav2Vec2 and Whisper. These models are pre-trained on extensive datasets of multilingual standard speech and are fine-tuned on the entire CapisciAMe corpus. This is of paramount importance for our work, as it allows us to recognize single voice commands and short sentences (composed of a combination of isolated words) uttered by speakers with atypical speech and dysarthria.

As discussed in the rest of the chapter, our contribution is not to introduce novel deep learning architectures for impaired speech recognition. Instead, we utilize open-source architectures (available in the literature) trained from scratch or fine-tuned on our private corpus of disordered speech in Italian. It is worth noting that both methodologies are not designed for continuous speech recognition, which is outside the scope of our work.

5.1 Isolated word recognition task with CNN

The research presented in this dissertation primarily focuses on isolated word recognition, aiming to identify specific predetermined keywords (single words) pronounced by selected

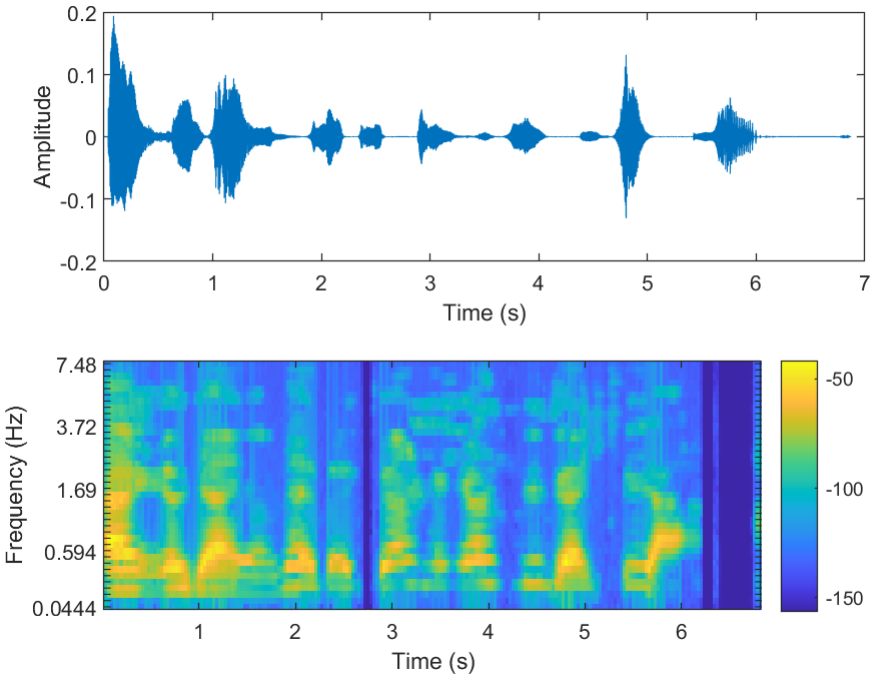


Figure 5.1: Example of a waveform (top) and voicegram (bottom) extracted from a dysarthric speech recording

Italian-speaking users with atypical speech and dysarthria. This approach is purely speaker-dependent, utilizing a small ASR vocabulary. Initially, this vocabulary consisted of thirteen (13) elements (*alto, basso, destra, sinistra, avanti, indietro, okay, chiudi, apri, tappa, entrata, uscita, volume*), including basic terms used to interact vocally with a computer by simulating mouse actions. To achieve a high-performance speech recognition system, our initial approach leverages multi-layer CNN architectures to address the challenges of phoneme alternation and imprecision in disordered speech. Instead of relying on a standard acoustic modeling component to identify the sequence of phonemes in voice commands, we employ a solution that visually presents speech features. The proposed method is based on voicegram analysis, which involves the graphical visualization of spectrum frequencies within the signals and their dynamics over the recording time, essentially a spectrogram specifically used for analyzing speech signals. This visual representation of the spectrum of frequencies in a signal as it varies with time provides a three-dimensional view of a sound signal, with time on the horizontal axis, frequency on the vertical axis, and amplitude represented by the intensity or color of each point in the image, as reported in Figure 5.1. This analysis can help us extract correlations within voicegrams (spectrograms) generated by different speakers with speech impairments who utter the same speech command.

More specifically, we have used the cnn-trad-fpool3 architecture, a baseline architecture for keyword spotting presented by Sainath and Parada from Google [114]. This model is composed of several layers that process the input audio features to produce a classification output. The key components of this convolutional architecture are as follows:

1. Input Layer: The input to the model is typically a set of Mel-Frequency Cepstral Coefficients (MFCCs), which are a representation of the audio signal. MFCCs are widely used in speech and audio processing because they capture the important features of the sound.

2. Convolutional Layers: The model includes two convolutional layers. Each convolutional layer applies a set of filters to the input data to detect various features. The first convolutional layer has 20 filters of size 8x8, and the second convolutional layer has 64 filters of size 1x3. These layers help in capturing the temporal and spectral features of the speech signal.

3. Pooling Layer: Following the convolutional layers, a max-pooling layer is applied. In the cnn-trad-fpool3 model, the pooling operation is performed with a filter size of 1x3. Pooling reduces the dimensionality of the data, making the model more efficient and less prone to overfitting.

4. Fully Connected Layers: After the convolutional and pooling layers, the model includes a fully connected (dense) layer. This layer combines all the features extracted by the convolutional layers and produces a high-level representation of the input data.

5. Output Layer: The final layer is a softmax layer that outputs the probabilities of each class (keyword). The class with the highest probability is chosen as the detected keyword.

Details on the internal structure of this CNN model are provided in Table 5.1, while a visual representation is presented in Figure 5.2. From a technological standpoint, a key benefit of this convolutional architecture is its deployability on smart computing devices, such as single-board computers or smartphones, where memory footprint can be limited. To this end, the artificial neural network contains two convolutional layers, unlike the state-of-the-art very deep and large CNNs, and there is also a limitation in the overall number of parameters. The cnn-trad-fpool3 architecture has been implemented into a TensorFlow graph, performing multiclass or multinomial classification tasks, i.e., classifying instances into one of three or more classes. The process leverages knowledge from the field of image classification by converting audio, a one-dimensional continuous signal across time, into a

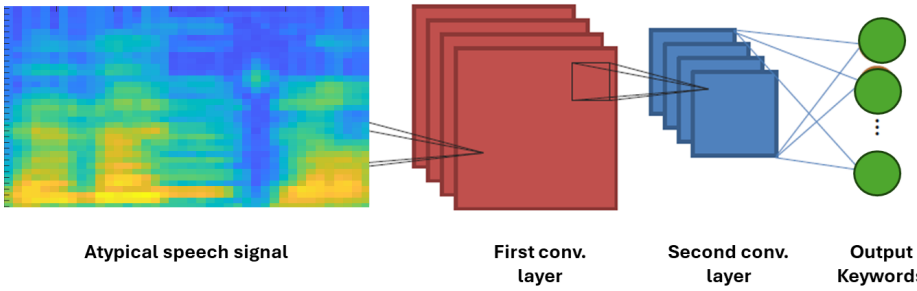


Figure 5.2: The two-layers CNN model used for isolated word recognition tasks

Type	**Ht.**	**Wd.**	**Dept**	**Stride Ht.**	**Stride Wd.**	**Par.**
Conv1	60	8	64	1	3	30.72k
Conv2	30	8	64	1	1	15.36k
Lin	-	-	32	-	-	65.5k
DNN	-	-	128	-	-	4.1k
Softmax	-	-	15	-	-	0.7k

Table 5.1: Details of the cnn-trad-fpool3 model

two-dimensional spatial problem. TensorFlow's utilities address this by defining a window of time that our speech commands should fit into and converting the audio signal within that window into an image. This conversion is done by grouping the incoming audio samples into short segments, just a few milliseconds long, and calculating the strength of the frequencies across a set of bands. Each set of frequency strengths from a segment is treated as a vector of numbers, and those vectors are arranged in time to form a two-dimensional array. We use Mel-Frequency Cepstral Coefficients as features [115], considering adjustable time windows [23].

During our research, we investigated the effectiveness of the aforementioned convolutional model on dysarthric speech in Italian, specifically for isolated word recognition tasks, as documented in published papers [107, 68, 67]. In particular, the contribution in:

- Mulfari, D., La Placa, D., Rovito, C., Celesti, A., & Villari, M. (2022). Deep learning applications in telerehabilitation speech therapy scenarios. Computers in Biology and Medicine, 148, 105864. DOI: 10.1016/j.compbiomed.2022.105864

proposed the utilization of the cnn-trad-fpool3 architecture, trained from scratch on a reduced partition of the CapisciAMe corpus, which included 5 hours of annotated dysarthric speech (at the time of that article). The aim was to exploit the impaired ASR within speech therapy and tele-rehabilitation scenarios. In these contexts, the repetition of relevant terms or sequences of keywords may facilitate users with dysarthria in performing tailored articulation exercises under the supervision of a speech-language pathologist, who may also operate

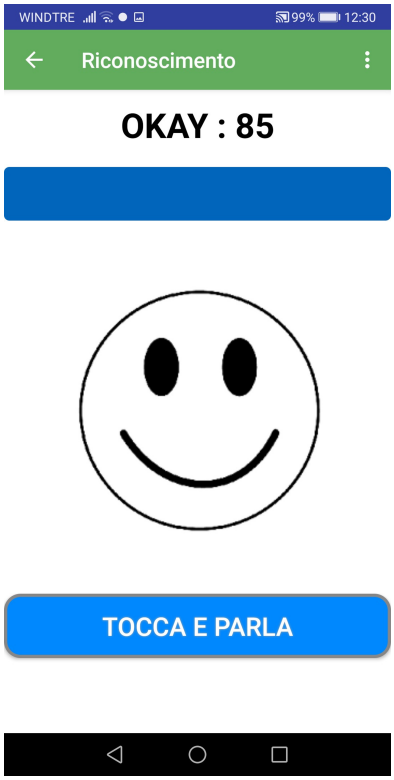


Figure 5.3: The app recognizes successfully the keyword "okay" spoken by the user.

remotely. To this end, the on-edge speech model inference process was deployed on a special version of the CapisciAMe app, along with adequate audio and video feedback needed to stimulate speech production from end users during speech therapy sessions. An example of positive video feedback is shown in Figure 5.3. With the collaboration of six male adults who have varying levels of dysarthria (mild, moderate, severe) due to different conditions (including cerebral palsy, neurodegenerative disease, and traumatic brain injury), we evaluated the performance of the local ASR tool (deployed on an Android device) in terms of accuracy for isolated word recognition. Using a 13-label dictionary, our best results showed a mean accuracy of over 90% across all dysarthria conditions. When the model was trained to recognize specific labels spoken by an individual with dysarthria, accuracy increased to 98%. This indicates that personalized training significantly enhances the ASR tool's performance, making it more reliable for individual users. As demonstrated in the article [68], these users can contribute to the speech model training by recording only 20 examples per keyword to obtain an overall word recognition accuracy greater than 90%.

Such positive results in Word Recognition Accuracy (WRA) have also been confirmed by increasing the size of our atypical speech dataset. Specifically, in the conference proceeding:

- Mulfari, D., Celesti, A., & Villari, M. (2022, May). Exploring AI-based Speaker De-

pendent Methods in Dysarthric Speech Recognition. In 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid) (pp. 958-964). IEEE. DOI: 10.1109/CCGrid54584.2022.00117

we demonstrated that enriching our 13-label corpus (specifically, up to 15 hours of transcribed dysarthric speech data) does not negatively impact the accuracy performance of our keyword recognizer for atypical voices. In fact, as shown in that paper, the WRA percentage remains high (around 96%) based on experimental evaluations conducted with the collaborative effort of thirteen Italian-speaking individuals with dysarthria linked to neuromotor disabilities, such as spastic quadriplegia. Furthermore, the study highlighted that a minimal number of repetitions (for example, 5) enables our CNN model to adapt to an unknown dysarthric speaker. Our quantitative analysis showed that, in cases of severe dysarthria, a slight increase in our training dataset (in the order of 65 speech examples totally) led to an improvement in WRA (approximately 11%). This represents a crucial aspect of our research, and more detailed investigations can be found in [107].

Similar results are described in a more recent paper [67], where the ASR dictionary was increased up to 54 terms. The research demonstrates that increasing the size of the atypical speech dataset does not negatively impact the accuracy of the ASR tool. This suggests that the system can handle larger datasets while maintaining high performance, which is crucial for developing robust speech recognition systems for individuals with dysarthria. At the same time, the ability of the ASR tool to maintain high accuracy with an expanded dataset indicates its scalability. This means the system can be adapted to recognize a broader range of speech patterns and commands, making it more versatile and useful in various real-world applications.

5.2 Sequence-to-sequence models for speech recognition

In the realm of automatic speech recognition (ASR), sequence-to-sequence (Seq2seq) architectures represent a significant advancement in how machines interpret human speech. This technology employs a sophisticated deep learning framework known as the encoder-decoder model, which processes and translates spoken language into text by mapping sequences of speech frames directly to sequences of characters. Unlike traditional ASR systems, which require separate training for acoustic, language, and pronunciation models, Seq2seq ASR integrates these components into a unified learning process. This integration simplifies the training architecture and has the potential to enhance the accuracy of speech recognition.

More specifically, encoder-decoder architectures are a type of artificial neural network designed for sequence-to-sequence tasks, such as machine translation and text summarization. These architectures consist of two main components, namely encoder and decoder. The encoder processes the input sequence and transforms it into a fixed-size context vector. This vector captures the essential information from the input sequence. The decoder takes the context vector and generates the output sequence, one element at a time, based on the encoded information.

As highlighted in recent literature [116], the use of encoder-decoder architectures in ASR offers several key advantages over convolutional neural networks structures, such as those described in the previous section. Encoder-decoder architectures are inherently designed to handle variable-length input and output sequences, making them well-suited for tasks like speech recognition and machine translation. In contrast, CNNs typically require fixed-size inputs and outputs, which can be a limitation for sequence-to-sequence tasks. The encoder-decoder model, especially when combined with attention mechanisms, can capture and utilize contextual information more effectively. This allows the model to focus on relevant parts of the input sequence when generating each part of the output sequence. CNNs, while powerful for spatial data, may not capture long-range dependencies as effectively. Seq2seq models integrate the acoustic, language, and pronunciation components into a single learning process. This unified approach can simplify the training pipeline and potentially improve the overall performance of the ASR system. CNNs, on the other hand, often require separate models for different components, which can complicate the training and integration process. Encoder-decoder architectures, particularly those based on transformers, offer greater flexibility and adaptability. They can be easily extended to incorporate additional features, such as attention mechanisms, which enhance their ability to model complex dependencies in the data. CNNs, while highly effective for image and spatial data, may require more modifications to achieve similar flexibility for sequence tasks. Encoder-decoder models, have shown superior performance on sequential data tasks compared to traditional CNNs. This is due to their ability to model temporal dependencies and relationships within the data more effectively.

As a consequence, while CNNs are highly effective for tasks involving spatial data, encoder-decoder architectures offer significant advantages for sequence-to-sequence tasks, such as ASR, due to their ability to handle variable-length sequences, capture contextual information, and integrate multiple components into a unified learning process.

Recent years have seen significant exploration into the impact of Seq2seq architectures and Transformer-based ASR solutions on the automated recognition of disordered speech. For example, Shahamiri et al. [30] evaluated the development of a Dysarthric Speech Transformer (DST), a speaker-adaptive, end-to-end dysarthric ASR system that employs trans-

former architectures with multiple encoder and decoder modules. This methodology was successfully tested on English using the UAspeech database; however, it remains limited to English and is specifically designed for dysarthria.

To address these limitations, we propose to employ the same sequence-to-sequence architecture used in contemporary automatic speech recognition systems. Specifically, we utilize pre-trained speech models. These models, which have emerged from recent advancements in deep learning, are becoming increasingly popular in various areas of speech technology. Pre-trained models are particularly attractive in fields such as impaired speech recognition, where speech datasets are typically small. They enable the use of deep neural networks that are initially trained on large datasets of standard speech and later adapted for areas with limited training data. In this context, our research exploits fine-tuning, namely a process where a pre-trained model is further trained on a smaller, task-specific dataset. This approach leverages the general knowledge the model has already acquired from a large, diverse dataset and adapts it to perform well on a specific task with limited data. In the context of ASR, fine-tuning involves adjusting the parameters of a pre-trained speech recognition model to improve its performance on recognizing speech from individuals with dysarthria and other speech impairments, as described in our recent articles [110, 109]. Specifically, during our PhD research, two distinct pre-trained, state-of-the-art ASR architectures have been investigated:

- Wav2Vec2 by Meta AI.
- Whisper by OpenAI.

The approach we propose for disordered speech recognition is particularly beneficial for languages and dialects with scarce labeled datasets nowadays. Despite there being over 7,000 languages worldwide, most lack the extensive training resources available for major languages like English. Statistics indicate that around 40% of these languages are endangered [117]. The limited availability of transcribed data for these low-resource languages hinders the effective training of large neural networks, resulting in subpar performance and limited practical applications. Consequently, automatic speech recognition for low-resource languages has increasingly become a focal point of global research [118]. Pre-trained models address this challenge by leveraging large amounts of unlabeled data during their initial training phase. For instance, the Wav2Vec2 architecture uses self-supervised learning to learn speech representations from vast quantities of raw audio. This means the model can capture a wide range of phonetic and acoustic patterns without needing labeled data. Once pre-trained, these models can be fine-tuned on smaller, task-specific datasets, making them

highly adaptable to new languages and dialects with limited labeled data, such as in the presence of dysarthria and other speech impairments.

OpenAI Whisper, on the other hand, employs weakly supervised learning and is designed to handle multiple languages and tasks. This makes it particularly versatile for low-resource languages. By fine-tuning on even small amounts of labeled data, Whisper can effectively adapt to the specific characteristics of these languages and dialects, improving recognition accuracy and usability. This approach not only enhances the performance of ASR systems in underrepresented languages but also democratizes access to speech technology. It enables the development of ASR solutions for communities that previously lacked the resources to build their own systems, thereby promoting linguistic diversity and inclusion in the digital age.

The rest of this section explores the used ASR architectures and their application on our private database of atypical speech in Italian.

5.3 Wav2Vec2 by Meta AI

Wav2Vec2 is a state-of-the-art model for automatic speech recognition developed by Meta AI. It leverages self-supervised learning to extract meaningful speech representations from raw audio data, significantly improving the performance of ASR systems, especially in scenarios with limited labeled data. As reported in the recent literature, Wav2Vec2 achieved state-of-the-art performance with 1.8% and 3.3% of Word Error Rate (WER) on the Librispeech corpus, i.e., a large-scale dataset of approximately 1,000 hours of read English standard speech, by considering the "clean" and "other" subsets, respectively. Additionally, the same architecture pre-trained on 53K hours of unlabeled data and only ten minutes of labeled data was able to achieve a WER in the range of 4.8-8.2%. Such results have opened many opportunities for ASR development with limited amounts of annotated speech data [119],

The architecture of Wav2Vec2 is composed of four main components: the feature encoder, the context network, the quantization module, and the contrastive loss. The feature encoder is the first component of the Wav2Vec2 architecture. It processes raw audio waveforms and converts them into latent speech representations. This encoder consists of several convolutional layers that capture local dependencies in the audio signal. The output of the feature encoder is a sequence of latent representations that serve as the input for the subsequent components of the model. Following the feature encoder, the context network is responsible for modeling long-range dependencies in the speech signal. This network is typically implemented using a Transformer architecture, which is well-suited for capturing complex dependencies over long sequences. The context network processes the latent rep-

resentations from the feature encoder and produces context-aware representations. These representations incorporate information from the entire input sequence, making them more informative for downstream tasks such as ASR. The quantization module represents one of the unique aspects of Wav2Vec2 architecture. This module discretizes the continuous latent representations into a finite set of discrete units. The quantization process involves mapping the continuous representations to the nearest codebook entries, which are learned during training. This step is crucial for the self-supervised learning objective, as it allows the model to learn useful speech representations without relying on labeled data. The training of Wav2Vec2 involves a contrastive loss, which is designed to maximize the similarity between the true latent representations and their quantized counterparts while minimizing the similarity with negative samples. During training, the model masks a portion of the latent representations and predicts the masked parts based on the context provided by the unmasked parts. This contrastive learning objective encourages the model to learn robust and informative representations of speech.

As depicted in the block diagram shown in Figure 5.4, the Wav2Vec2 model encodes raw audio waveforms (in our case, atypical speech signal), denoted as χ, into latent speech representations $z_1....z_T$ via a multi-layer convolutional feature encoder $f : \chi \rightarrow Z$. Such a latent representations fed a Transformer-masked network $g : Z \rightarrow C$. The Transformer network initially quantize the continuous representations, forming a discrete set of outputs $z_1....q_T$, which represent targets in the self-supervised learning objective. Those quantized representations are then contextualized using the attention blocks from the Transformer module, obtaining a set of discrete contextual representations $c_1....c_T$. The feature encoder is formed by seven convolutional blocks with 512 channels, strides of $\{5, 2, 2, 2, 2, 2, 2\}$ and kernel widths of $\{10, 3, 3, 3, 3, 2, 2\}$. The Transformer network comprises 24 blocks, with a model dimension of 1024, an inner dimension of 4096, and a total of 16 attention heads [86].

5.3.1 Implications for our project

Within the framework of our PhD research, the scientific contribution published in:

- Mulfari, D., Carnevale, L., Galletta, A., & Villari, M. (2023, May). Edge Computing Solutions Supporting Voice Recognition Services for Speakers with Dysarthria. In 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW) (pp. 231-236). IEEE.
  DOI: 10.1109/CCGridW59191.2023.00047

provides a comprehensive investigation into the application of the Wav2Vec2 architecture for automatic disordered speech recognition, using a significant subset of the CapisciAMe

database in Italian. Specifically, the study employed the base version of Wav2Vec2, pre-trained on raw audio data sampled at 16kHz, and fine-tuned on 41 hours of transcribed disordered speech in Italian. The ASR dictionary in this study was limited to 71 keywords, with the primary goal of developing a voice recognition system capable of identifying only these speech commands within impaired utterances.

In this context, it is important to note that when a voice signal is input, the proposed speech model produces a sequence of characters that may not correspond to any terms in our ASR dictionary. To address this issue, the study introduced the concept of character distance between two strings and employed a simple K-Nearest Neighbors (KNN) algorithm to select the keywords from our dictionary that minimize the Levenshtein distance from the generated ASR output string. The adoption of this similarity function is crucial as it enables our ASR to function as a speech classifier for dysarthric speech. Under these conditions, our experimental evaluation focused on studying the performance of our speech recognition system in terms of word recognition accuracy (WRA), which represents the percentage of correctly recognized words out of the total words in the reference (ground truth) testing dataset, as defined by the following expression:

$$\text{WRA} = \frac{\text{Correct predictions}}{\text{All predictions}} \times 100 \tag{5.1}$$

Specifically, quantitative experiments conducted on a custom testing dataset, arranged with the collaboration of sixteen speakers with various levels of dysarthria (mild, moderate, severe), highlighted an overall WRA of 97.4%, with an accuracy greater than 95% across different dysarthria conditions, and consequently, a WER of approximately 5%.

5.4 Whisper by OpenAI

Whisper is an advanced, state-of-the-art automatic speech recognition system developed by OpenAI. It employs an end-to-end approach, mapping input audio signals directly to output text captions without relying on intermediate representations or separate processing stages. Central to the Whisper architecture is an encoder-decoder Transformer model, which effectively captures long-range dependencies and contextual information, enabling the model to learn complex mappings between input audio features and corresponding textual transcriptions.

As illustrated in Figure 5.5, the input speech signal is resampled at 16 kHz and segmented into 30-second chunks. This data is then converted into an 80-channel log-magnitude Mel

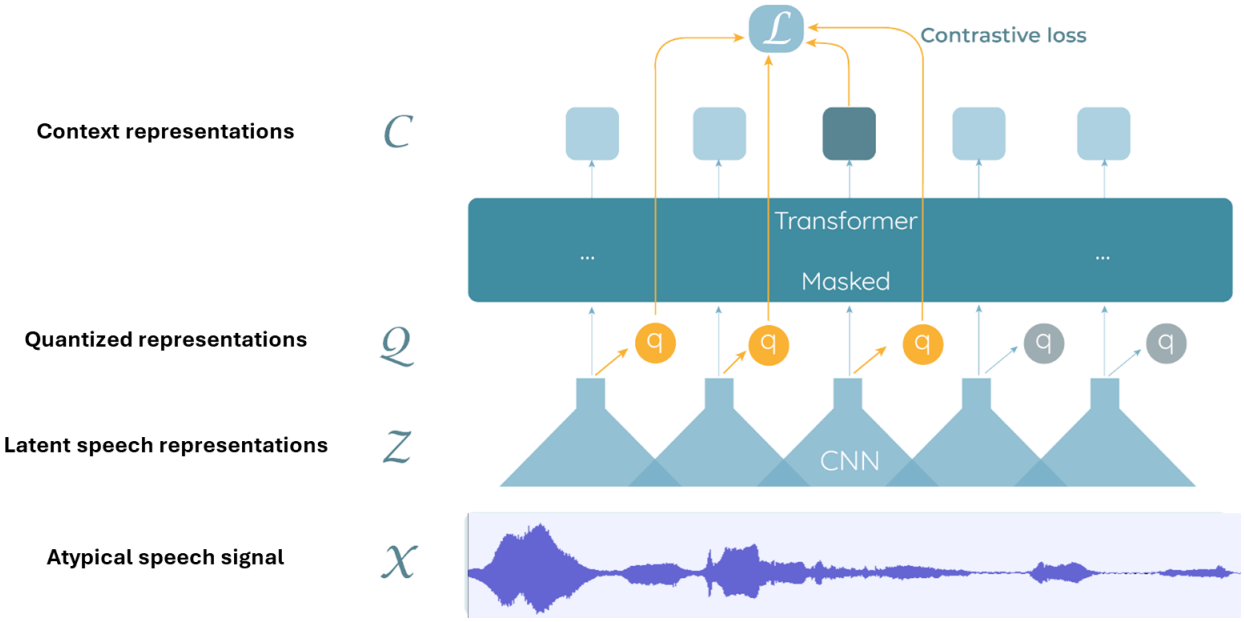


Figure 5.4: Structure of the Wav2Vec2 model by Meta AI used in the CapisciAMe project.

spectrogram using 25 ms windows with a 10 ms stride, and normalized to a range of [-1, 1] with near-zero mean to ensure consistent input for the model. The encoder consists of multiple layers of convolutional neural networks (CNNs) followed by transformer layers. The CNNs help in capturing local patterns in the spectrogram, such as phonemes and syllables, while the transformer layers capture the broader context and dependencies across the entire audio sequence. This combination allows the encoder to produce a rich, high-dimensional representation of the input audio.

The decoder in Whisper is responsible for generating the corresponding text from the encoded audio representation. It is trained to predict the next character or token in the sequence, given the previous tokens and the encoded audio features. The decoder uses a combination of attention mechanisms and feed-forward neural networks to generate the output text. A comprehensive explanation of the OpenAI's Whisper architecture is available in [56].

One of the key features of Whisper's decoder is its ability to handle multiple tasks simultaneously. It can perform language identification, phrase-level timestamping, multilingual speech transcription, and speech-to-text translation. This multitasking capability is achieved by intermixing special tokens in the training data that direct the model to perform specific tasks.

Whisper is currently trained on a massive dataset of 680,000 hours of multilingual and multitask supervised data collected from the web. This dataset includes a diverse range of

audio sources, such as podcasts, YouTube videos, and other publicly available speech recordings. The large and varied dataset helps Whisper achieve robustness to different accents, background noises, and technical languages. Specifically, the training process involves two main stages: pre-training and fine-tuning. During pre-training, the model learns general audio-to-text mappings from the large dataset. This stage helps the model develop a broad understanding of different languages and speech patterns. Fine-tuning is then performed on more specific datasets to adapt the model to particular tasks or domains, including the automated recognition of disordered speech, as shown in recent literature focused on ASR tasks in English [100].

By accommodating variations in speech quality and environmental noise, the model can adapt more effectively to the unique challenges posed by atypical speech patterns. This adaptability allows for targeted adjustments during fine-tuning processes, optimizing the model's performance to better capture the distinctive speech patterns and nuances characteristic of individuals with speech disabilities. Additionally, the Whisper model's resilience to such variations ensures that fine-tuning efforts remain effective across a range of real-world scenarios, ultimately leading to more accurate and reliable speech recognition outcomes for speakers with dysarthria and other classes of speech disorders.

At the time of writing, Whisper has several variants, each designed to cater to different needs and computational resources, as detailed below:

- *Whisper Tiny* has 39 million parameters and requires about 1GB of RAM. It is approximately 10 times faster than the large model and is available in an English-only version (tiny.en). This variant is suitable for environments with limited computational resources, offering fast inference times with reasonable accuracy.

- *Whisper Base* has 74 million parameters and also requires about 1GB of RAM. It provides a balance between speed and accuracy and is available in an English-only version (base.en). This variant is ideal for applications needing slightly more accuracy without a significant increase in computational requirements.

- *Whisper Small* has 244 million parameters and requires about 2GB of RAM. It offers a good trade-off between model size and performance, making it suitable for more demanding applications where higher accuracy is needed but computational resources are still a consideration.

- *Whisper Medium* has 769 million parameters and requires about 5GB of RAM. This variant is designed for high-accuracy applications, requiring more computational power

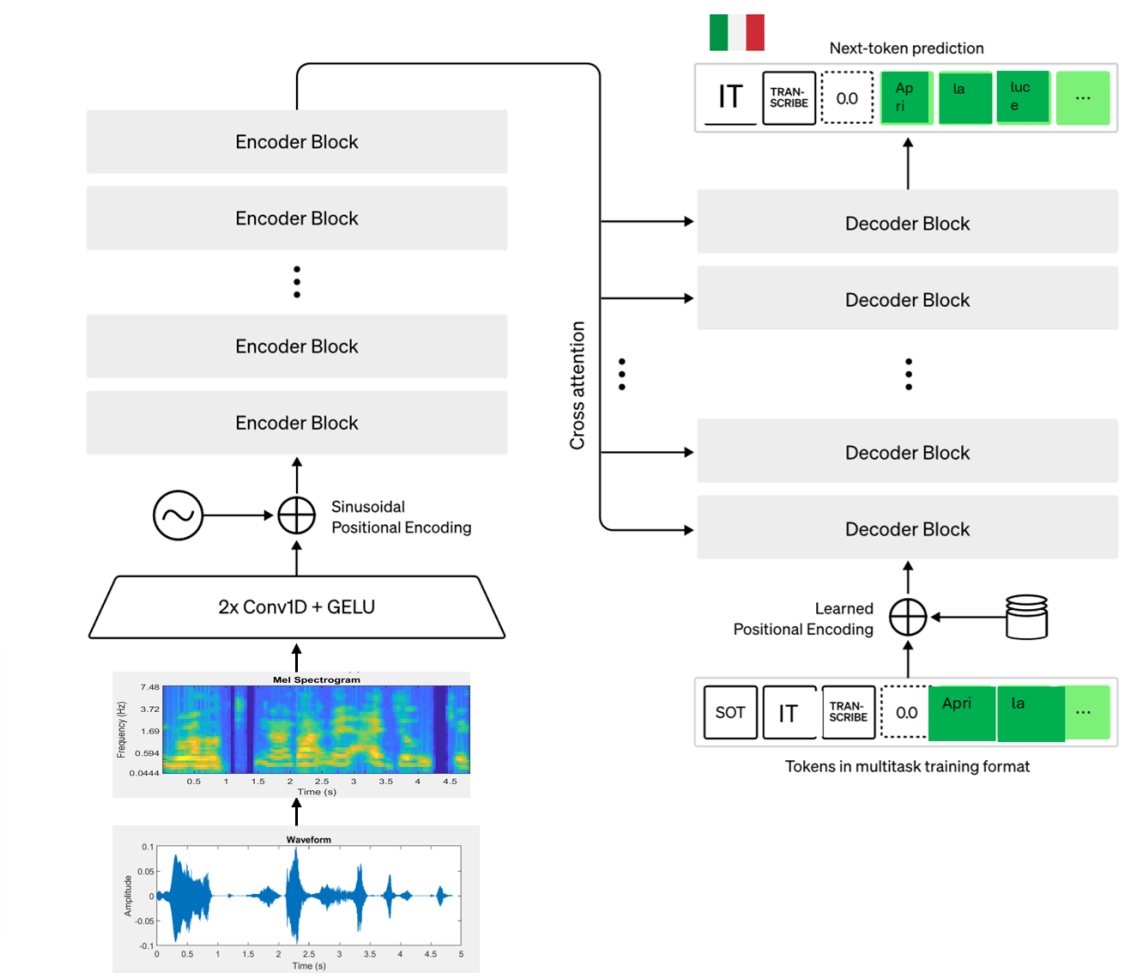


Figure 5.5: Structure of the Whisper model by OpenAI used in the CapisciAMe project

but delivering significantly better performance in terms of transcription accuracy and robustness to different accents and background noises.

- *Whisper Large* has 1.55 billion parameters and requires about 10GB of RAM. It is the most powerful variant in terms of accuracy and robustness, suitable for applications where the highest possible transcription quality is required, and computational resources are not a limiting factor.

- *Whisper Turbo* has 809 million parameters and requires about 6GB of VRAM. This recently released variant offers nearly the same capability as the Large model but with significant speed improvements, up to 8 times faster. This is achieved through fine-tuning a pruned version of the Large model, making it an excellent choice for high-performance applications that also require faster processing times.

5.4.1 Implications for our project

At this point, it is of interest to investigate the performance of the aforementioned speech model variants on atypical speech, in conjunction with our CapisciAMe corpus.

To this end, our study published in

- Mulfari, D., & Villari, M. (2024). A Voice User Interface on the Edge for People with Speech Impairments. Electronics, 13(7), 1389.
  DOI: https://doi.org/10.3390/electronics13071389

investigates the utilization of several Whisper variants, particularly Tiny, Base, and Small, with the specific aim of empowering portable Voice User Interfaces (VUIs) designed for speakers with speech impairments. Our objective was to run the speech model's inference on popular embedded systems and single-board computers, such as Raspberry Pi boards, which have an operating system and limited hardware resources. Therefore, on these IoT platforms, it is important to minimize RAM usage needed for on-edge speech model inference and to employ a software framework aimed at accelerating the speech recognition task. To meet these requirements, the analysis excluded the utilization of Medium, Large, and Turbo Whisper variants and used the "Whisper.cpp" framework [1], which provides a solution for executing inference tasks in the C++ programming language. Further details on the specific software implementation are provided in [109].

One of the main contributions of that research article is the investigation of the performance of the three selected OpenAI variants in isolated word recognition tasks, using a significant partition of the current version of our impaired speech dataset. Specifically, the accuracy experiments were performed on an archive composed of approximately 46 hours of labeled disordered speech in Italian, equivalent to more than 65,000 single speech examples belonging to a 79 keywords ASR dictionary. Evaluations conducted in collaboration with sixteen Italian-speaking individuals with dysarthria (who participated in speech model training) highlighted an overall word recognition accuracy of 95.9% and, consequently, a word error rate close to 4%. Table 5.2 summarizes the results from [109], grouped by the severity levels of dysarthria in the individuals who participated in our experiments.

From the above investigation, we found that the use of Whisper's small variant consistently led to improved accuracy in isolated word recognition compared to other configurations. This improvement was most pronounced in cases of severe speech impediment, where the small variant exhibited approximately a 5% WRA improvement over the base variant. When comparing the base and tiny Whisper configurations, we observed a more modest increase. Similar trends were evident for moderate and mild speech disorders, as summarized

[1] https://github.com/ggerganov/whisper.cpp

		OpenAI Whisper variants		
		Tiny	**Base**	**Small**
Dysarthria severity levels	**Severe**	82.9	85.2	90.3
	Moderate	88.6	91.8	93.1
	Mild	87.7	92.2	93.7

Table 5.2: Word recognition accuracy results (expressed in percentage) grouped by three different OpenAI Whisper variants

in the same table. It is worth noting that the base model faced challenges when dealing with severe dysarthria, particularly in instances where participants had a reduced number of personal speech samples in the training dataset. The reasons behind this abnormal performance remain unclear. We believe that the extremely acute dysarthria exhibited by these speakers, coupled with the scarcity of their speech samples, led to significant divergence from the speech features present in our CapisciAMe corpus of atypical voice samples in Italian.

It is important to note that our work based on OpenAI Whisper (Small variant) stands out compared to recent literature focused on atypical speech command recognition. The results we achieved surpass prior investigations: in [66], a 85% WRA was measured using a CNN model trained with only a subset of the current CapisciAMe speech collection. Similar results are reported in [67, 68], where the evaluation involved a small number of participants with dysarthria caused by infant cerebral palsy, using a different version of our private corpus of atypical speech.

As a consequence, a direct comparison based solely on numerical accuracy values might lack significance. Researchers have employed different corpora and methodologies to evaluate their ASR solutions, leading to variations in results. Factors such as the structure of speech databases, the chosen language, the input modality, and the varying levels of speech disability all play crucial roles, as motivated in a recent review [4].

5.5 Experimental evaluation

The results presented in the previous sections have highlighted the effectiveness of Wav2Vec2 and Whisper for automatic disordered speech recognition. Both state-of-the-art ASR architectures have been fine-tuned using different subsets of the CapisciAMe database for isolated word recognition tasks, where our analysis showed remarkable performance in terms of single word recognition accuracy.

In this section, we aim to take a step further by investigating the performance of these two architectures on the same testing dataset, which now includes both isolated words and

	Examples	Recordings Time	Words	Sentences
Training Set	77,736	57 h	67,541(49 h)	10,195(8h)
Testing Set	8,638	6 h	7,505(5h)	1,133 (1h)

Table 5.3: Details of the training and testing datasets used in our experiments

short sentences. Each short sentence is composed of a sequence of single words from our limited ASR dictionary. To conduct these experiments, a significant portion of the current CapisciAMe corpus has been utilized. This block encompasses a total of 63 hours of annotated disordered speech in Italian, divided into two separate datasets: a training dataset and a testing dataset, with no overlapping examples. Table 5.3 reports the details of the two corpora, built with the collaborative effort of 195 Italian-speaking individuals with dysarthria and other speech disorders. It is important to highlight that all the speech recordings have been verified to suppress disturbances and noise components, and to ensure accurate text transcription.

5.5.1 Experiments setup

The same training dataset content was used to fine-tune the Meta AI Wav2Vec2-base [2] and the OpenAI Whisper architectures. Specifically, we selected the Small variant of Whisper [3], considering the positive results discussed previously.

All the speech model training operations were conducted on the Artemis high-performance computing cluster at the University of Sydney. This cluster is designed for complex data analysis across various research areas, including , molecular biology, economics, mechanical engineering, and oceanography. Specifically, we used Python 3.7 and PyTorch 1.3 as the programming language and deep learning framework, respectively. The virtual environment also supported the NVIDIA CUDA 10.2 libraries. All deep learning processes were included in a job executed on a single computing node of the cluster, which featured a quad-core CPU, 128 GB of RAM, and one NVIDIA V100 GPU. In this virtual software environment, the Transformers library (version 4.26.1) was used to obtain the two speech models from the HuggingFace platform and to conduct all necessary fine-tuning operations according to the Transformers library guidelines [4]. Details on the used hyperparameters are presented in Table 5.4.

The learning curves of Wav2Vec2 and Whisper, fine-tuned on the same disordered speech dataset, are reported in Figures 5.6 and 5.7. These graphs provide a compelling comparison

[2] https://huggingface.co/facebook/wav2vec2-base
[3] https://huggingface.co/openai/whisper-small
[4] https://huggingface.co/

Hyperparameter	**Explanation**	**Wav2Vec2**	**Whisper**
`per_device_train_batch_size`	Number of training samples per GPU device per batch	16	16
`learning_rate`	Learning rate for the optimizer	1e-5	1.25e-5
`warmup_steps`	Number of steps to gradually increase the learning rate at the start	1000	1000
`num_train_epochs`	Total number of training epochs	30	3
`gradient_checkpointing`	Saves memory by not storing intermediate activations during backpropagation	True	True
`fp16`	Uses 16-bit floating point precision to speed up training and reduce memory usage	True	True
`evaluation_strategy`	Strategy for evaluation during training (e.g., steps, epochs)	Steps	Steps
`per_device_eval_batch_size`	Number of evaluation samples per GPU device per batch	8	8
`save_steps`	Number of steps between saving model checkpoints	500	500
`eval_steps`	Number of steps between evaluations	500	500
`logging_steps`	Number of steps between logging training metrics	500	25

Table 5.4: Hyperparameters used to fine-tune Wav2Vec2 and Whisper

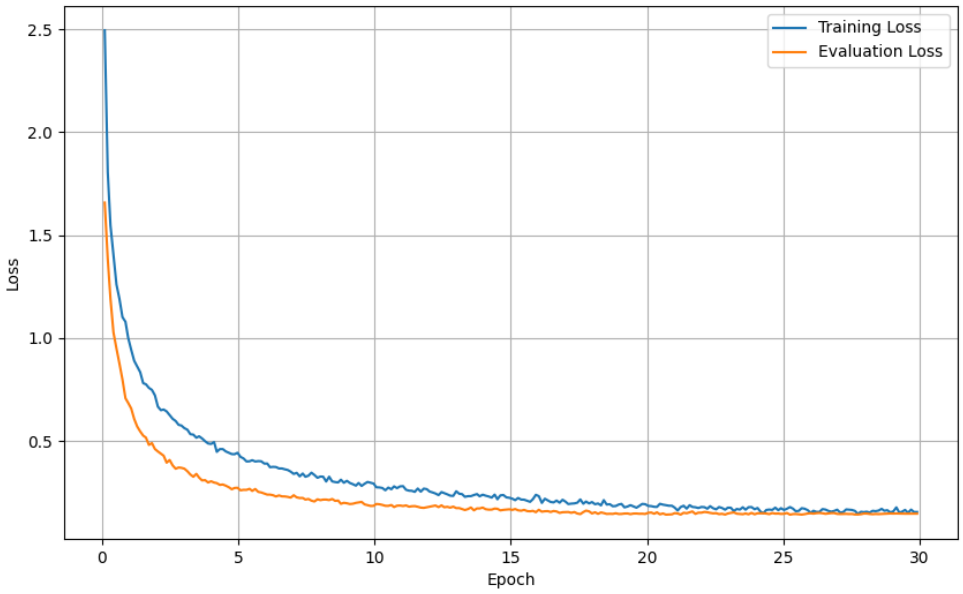


Figure 5.6: Meta AI Wav2Vec2 fine-tuning process: learning curve

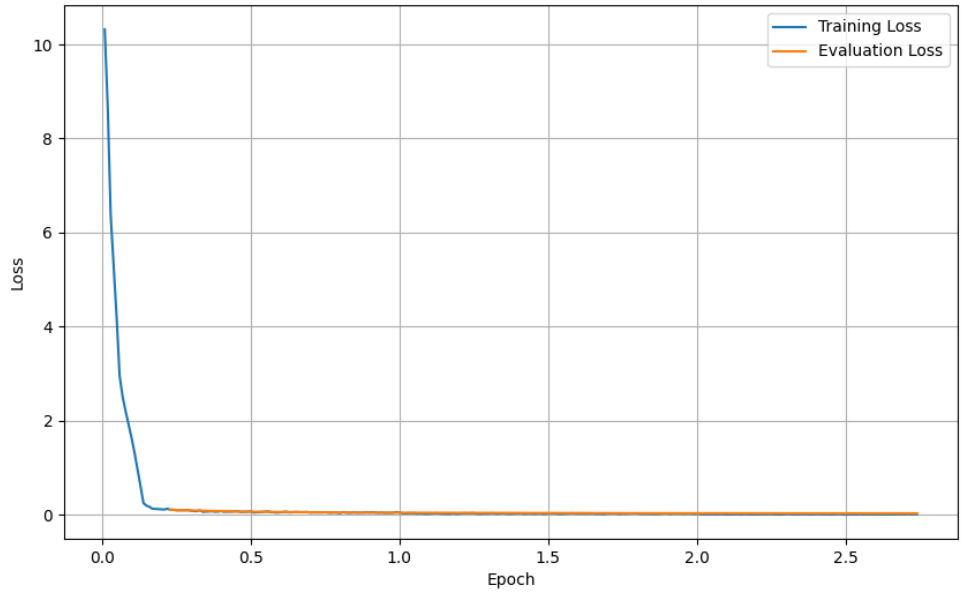


Figure 5.7: OpenAI Whisper fine-tuning process: learning curve

of the two ASR architectures' behaviors during training and their ability to adapt to our dataset. Both models, state-of-the-art in automatic speech recognition, exhibit distinct characteristics in terms of convergence speed, computational efficiency, and generalization capability.

For Wav2Vec2, the learning curve on the dysarthric speech dataset shows a gradual reduction in both training and evaluation losses over 30 epochs. Initially, the training loss is high, indicating the model's unfamiliarity with the dataset, but it steadily declines as the model learns. A similar trend is observed for the evaluation loss, which remains consistently lower than the training loss throughout the fine-tuning process. This behavior suggests that Wav2Vec2 generalizes well to the validation data, avoiding overfitting even with extended training. The evaluation loss stabilizes after approximately 15 epochs, with only marginal improvements beyond this point, indicating a plateau where additional epochs yield diminishing returns. However, the extended training period highlights the computational demands of Wav2Vec2. Its slow but steady convergence suggests that while it can capture the nuanced patterns of dysarthric speech, it requires substantial training time to achieve optimal performance.

In contrast, Whisper's learning curve shows rapid convergence, with both training and evaluation losses stabilizing within the first epoch. The near-zero values for loss suggest that Whisper is highly effective at fine-tuning with minimal training, leveraging its extensive pre-training to quickly adapt to the dysarthric speech dataset. This efficiency is particularly notable given the dataset's complexity. The near-identical trends in training and evaluation loss indicate that the model fits the data well, but this could also signal potential overfitting. While Whisper's architecture allows for rapid adaptation, its performance must be validated on an independent test set to ensure that the observed low loss values indicate true generalization rather than over-reliance on the training and validation data.

Whisper's ability to converge quickly underscores its robustness as a pre-trained model optimized for diverse speech tasks. It is particularly advantageous for scenarios where computational resources are limited, and rapid deployment is necessary. However, this efficiency might come at the cost of missing finer dataset-specific patterns that models like Wav2Vec2 can learn with extended training. For dysarthric speech, which often requires a more nuanced understanding of speech irregularities, this could be a limitation depending on the specific application requirements.

When comparing the learning curves and training behaviors of Wav2Vec2 and Whisper, several key differences emerge. First, the convergence speed of the two models is starkly different. Wav2Vec2's loss decreases gradually over 30 epochs, requiring significant computational resources and training time. Whisper, on the other hand, converges within a single

epoch, making it much more computationally efficient. This rapid convergence suggests that Whisper can be fine-tuned quickly, a critical factor in scenarios where time and resources are constrained, especially in terms of GPU-enabled resources.

Second, the trends in evaluation loss offer insights into the models' generalization capabilities. Wav2Vec2's evaluation loss remains consistently lower than its training loss, indicating effective regularization and an ability to generalize well to unseen data. Whisper's evaluation loss, however, closely mirrors its training loss, which could suggest overfitting despite the low final loss values. This difference underscores the need for careful validation of Whisper's performance on independent test data, particularly for dysarthric speech, where overfitting could lead to poor real-world performance.

Third, the magnitude of loss at convergence provides additional perspective. Whisper achieves near-zero loss values within the first epoch, reflecting its robustness as a pre-trained model. In contrast, Wav2Vec2 requires significantly more epochs to reach low loss values, highlighting its dependence on extensive fine-tuning. This distinction reflects the different training paradigms of the two models: Whisper's reliance on large-scale pre-training for general-purpose adaptability versus Wav2Vec2's focus on learning dataset-specific features through extended training.

Upon completion of the speech model training operations, the resulting model checkpoints were generated and transferred to our workstation for subsequent inference analysis. This desktop environment, equipped with 32 GB of RAM and a single NVIDIA RTX3060 GPU, was used to evaluate inference results using the same testing dataset.

5.5.2 Performance metrics

The same testing dataset, consisting of six hours of transcribed disordered speech in Italian, has been utilized to investigate the performance of the previously fine-tuned ASR model in terms of recognition accuracy. Since the testing dataset includes both sentences and single words, we chose the word error rate (WER) as main performance index.

The WER is a common metric used to evaluate the performance of ASR systems and it represents the cost of restoring the output word sequence to the original input sequence [120]. It measures the difference between a reference transcription (the correct text) and the hypothesis transcription (the text generated by the ASR system). The WER is calculated as the sum of the number of substitutions, deletions, and insertions needed to transform the hypothesis into the reference, divided by the total number of words in the reference, and then multiplied by 100 to express it as a percentage.

The WER is calculated using the following formula:

$$\text{WER}(\%) = \left(\frac{S + D + I}{N}\right) \times 100 \tag{5.2}$$

where:

- S is the number of substitutions (incorrect words in the hypothesis transcription),
- D is the number of deletions (missing words in the hypothesis transcription),
- I is the number of insertions (extra words the hypothesis transcription),
- N is the total number of words in the reference transcription.

It is important to note that WER may exceed 100% because the calculation includes insertions, which are extra words that the ASR system incorrectly adds to the transcription. When the number of insertions ((I)) is very high, it can significantly increase the numerator (the sum of substitutions, deletions, and insertions). If this sum exceeds the total number of words in the reference ((N)), the resulting WER value will be greater than 100

WER is crucial for comparing different ASR systems and tracking improvements over time. It helps developers understand how well their models are performing and identify areas for improvement. However, WER has limitations, such as not accounting for the severity of errors (e.g., missing a critical word versus a minor one) and not being sensitive to word order. Despite these limitations, WER remains a widely used and valuable metric in the field of speech recognition. As such, we adopt WER as major performance metric for our study. In addition, we consider Match Error Rate (MER) as a secondary metric for evaluating the performance of our ASR tool. The MER value represents the percentage of words that were incorrectly predicted and inserted the hypothesis transcription. It is defined by the following equation:

$$\text{MER}(\%) = \left(\frac{S + D + I}{N + I}\right) \times 100 \tag{5.3}$$

5.5.3 Results and discussion

To conduct our quantitative analysis, all the recognition results have been organized into a suitable CSV file. As sequence-to-sequence models, both ASR architectures used (Whisper by OpenAI and Wav2Vec2 by Meta AI) convert an input acoustic speech waveform into a sequence of characters forming words or sentences. Since we have the correct transcription for each impaired speech recording in our database, we can assess the word error rate of our ASR system.

This calculation employs the WER implementation provided by the "jiwer" Python library [5], which contains various tools designed to analyze the performance of an automatic speech recognition system.

```
import sys
from jiwer import wer
import pandas as pd
df = pd.read_csv(sys.argv[1])
quanti = len(df['Word'])
r=[]
h=[]
for i in range(0,quanti):
    r.append(str(df['Word'][i]))
    h.append(str(df['Word_I'][i]))
c=wer(reference=r,hypothesis=h)*100
c=round(c,1)
print(c)
```

Listing 5.1: WER calculation script

Listing 5.1 report the script utilized to calculate the percentage of WER from our CSV file passed as an argument. A similar code was used to measure the percentage of MER. We have estimated the overall metrics of the two distinct ASR architectures by concatenating all the hypotheses with all the references. This quantitative investigation has led us to the following results when both Whisper and Wav2Vec2 were fine-tuned on the same block of our private disordered speech corpus in Italian:

- a WER of 3.5% using Whisper, with a MER of 3.5%;
- a WER of 5.8% using Wav2Vec2, with a MER of 5.8%.

It is worth noting that, in this specific case, when rounding to one decimal place, the WER and MER values are equal. This indicates that the number of insertions in the ASR-generated transcription is zero.

A visual comparison between these results is depicted in Figure 5.8. The results highlight that Whisper, particularly its "Small" variant, made fewer errors in transcribing the disordered speech compared to the "Base" configuration of Wav2Vec2. This suggests that OpenAI's ASR architecture is more effective at handling the nuances and irregularities present in the disordered speech dataset. The lower WER reflects Whisper's ability to generalize

[5] https://jitsi.github.io/jiwer/

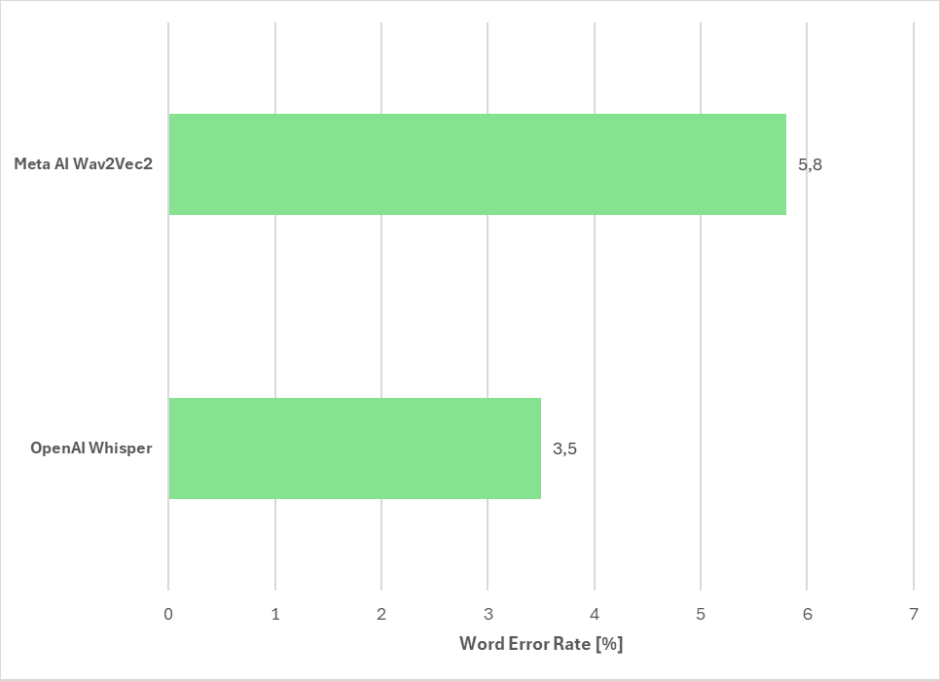


Figure 5.8: WER results (expressed as a percentage) of Whisper and Wav2Vec2 fine-tuned on disordered speech

well to the specific characteristics of the dataset, likely due to its extensive pre-training on a diverse set of speech data. In contrast, Wav2Vec2 achieved a higher WER than Whisper. This indicates that Wav2Vec2 made more errors in transcribing the disordered speech. While Wav2Vec2 is still a strong performer, the higher WER suggests that it may not be as well-suited to the specific challenges posed by the disordered speech dataset as Whisper. This could be due to differences in the pre-training data and the model architecture. Comparing the two models, Whisper's lower WER suggests that it is more efficient and adaptable to the disordered speech dataset. This could be attributed to its training on a large, diverse dataset that includes various speech patterns and accents, making it more robust in handling atypical speech. The results highlight the importance of the pre-training phase. Whisper's extensive pre-training on multilingual and multitask data likely contributed to its superior performance. In contrast, Wav2Vec2, while effective, may require more specialized fine-tuning or additional data to match Whisper's performance on this specific task. Depending on the application, Whisper might be the preferred choice for tasks involving disordered speech due to its lower error rate. However, Wav2Vec2 could still be valuable, especially if further fine-tuning or additional training data can help reduce its WER. Overall, these results suggest that Whisper is currently better suited for recognizing disordered speech in Italian, but both models have their strengths and potential areas for improvement.

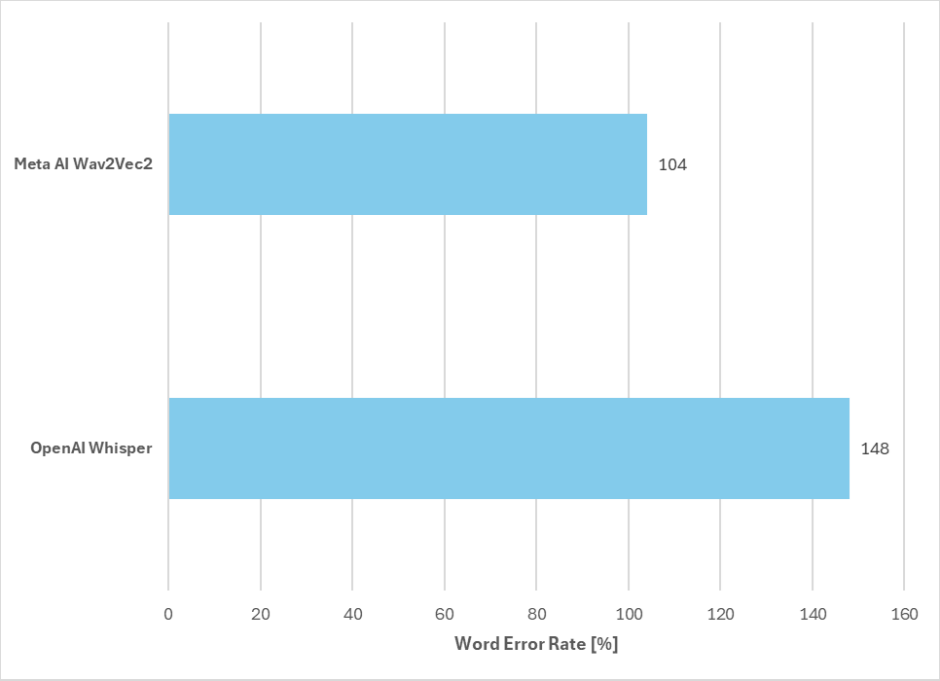


Figure 5.9: WER results (expressed as a percentage) of Whisper and Wav2Vec2: no fine-tuning applied

Furthermore, to investigate the added value of fine-tuning the aforementioned state-of-the-art ASR architectures on impaired speech, we analyzed the condition where no fine-tuning was performed. Using the same testing dataset of impaired speech, i.e., a subset of the current CapisciAMe database, we observed a significant increase in WER. As shown in Figure 5.9, under these conditions, Whisper reported a WER of 148% (with a MER of 95%) and Wav2Vec2 a WER of 104% (with a MER of 91%). These results indicate that these speech models are not currently designed to recognize atypical speech and confirm the poor performance of contemporary ASR systems for speakers with speech disorders.

5.5.4 Conclusion

The experimental evaluation presented in this chapter highlights the critical importance of fine-tuning state-of-the-art ASR architectures, such as Whisper and Wav2Vec2, on our impaired speech dataset in Italian, which encompasses a total of 63 hours of transcribed voice recordings (within the framework of the proposed experiments). Without fine-tuning, both models exhibited significantly higher word error rates (WER), with Whisper reporting a WER of 148% and Wav2Vec2 a WER of 104%. These elevated error rates underscore the current limitations of these models in recognizing atypical speech patterns and reveal a bias in the multilingual databases used to train current ASR systems, where non-standard speech

patterns are underrepresented.

Fine-tuning on a specific dataset of disordered speech dramatically improved performance, reducing the WER to 3.5% for Whisper and 5.8% for Wav2Vec2. This improvement demonstrates the added value of tailoring ASR systems to the unique characteristics of impaired speech. Whisper's superior performance, particularly its "Small" variant, suggests that its extensive pre-training on diverse and multilingual data makes it more adaptable to the nuances of disordered speech.

However, the higher WER observed in the absence of fine-tuning indicates that contemporary ASR systems are not inherently designed to handle atypical speech. This finding confirms the need for specialized training and adaptation to improve ASR performance for speakers with speech disorders.

In conclusion, while Whisper currently shows greater efficiency and adaptability for recognizing disordered speech, both Whisper and Wav2Vec2 have potential areas for improvement. Future research should focus on enhancing the robustness of these models through more extensive pre-training and fine-tuning, as well as exploring additional data sources to better accommodate the variability in impaired speech.

5.6 Summary

This chapter presented deep learning approaches to impaired speech recognition. Section 5.2 investigated the utilization of pre-trained state-of-the-art encoder-decoder architectures, namely Wav2Vec2 and Whisper, which have been fine-tuned on the CapisciAMe database. Experimental evaluations demonstrated the effectiveness of the proposed approaches in terms of word error rate, as detailed in Section 5.5.

Chapter 6

Speech recognition applications

This chapter addresses the third pillar of our digital ecosystem designed to support automatic speech recognition services for individuals with dysarthria and other speech impairments. It focuses on the development of several application prototypes aimed at demonstrating the utilization of our voice recognition engine in real-world contexts involving persons with neuromotor and speech disabilities. Our proposal falls within the framework of digital assistive technology, which refers to electronic devices, software, or systems that assist individuals with disabilities in performing functions that might otherwise be difficult or impossible [121]. Specifically, we focus on the field of speech interaction with computers and smart devices.

At the core of this work is the deployment of our speech recognition engine as a cloud-based service, enabling the real-time transcription of disordered speech in Italian. In the following sections, we first describe the proposed cloud-based speech recognition services for disordered speech transcription and then present a review of the prototype application scenarios.

6.1 Disordered speech transcription services

Speech transcription services convert spoken language into written text. To support our application scenarios, we leverage the findings of the previous chapter to implement custom automated services that apply our signal processing and deep learning algorithms. These services provide a textual response (sequences of characters) in response to the submission of an acoustic speech waveform containing disordered speech in Italian.

This resource has been deployed as a cloud-based service, enabling the process illustrated in Figure 6.1. It is important to highlight that our transcription service includes the software pipeline for speech signal enhancement presented in Section 4.2. In this way, the speech-

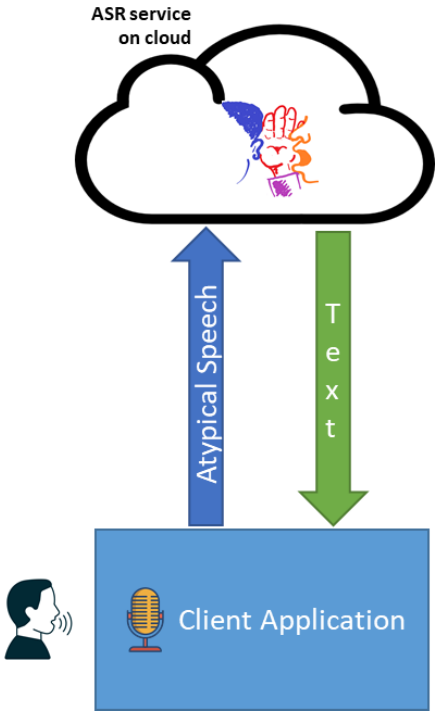


Figure 6.1: Block diagram of the transcription services

to-text conversion is deployed as an on-demand cloud service, playing a crucial role in our technological ecosystem since it provides software developers with a convenient way to integrate voice user interfaces into their custom applications. The availability of these ASR tools can significantly support other studies exploiting automatic disordered speech recognition in diverse fields, such as human-computer interaction, interpersonal communication, e-health, tele-rehabilitation, and speech therapy scenarios. In these contexts, the feedback received from a tailored speech recognition service can help individuals with dysarthria perform specific articulation exercises, providing therapists with valuable insights and tools to monitor and assist patients' progress. Further details on this application can be found in our study published in [68].

Additionally, another context lies in the field of human-robot interaction, where the utilization of our speech recognition services can empower voice interaction with assistive robots, making them accessible to people with dysarthria, as demonstrated by our recent collaboration with researchers at the "Campus Bio-Medico" University of Rome.

Therefore our speech transcription services are designed to power voice user interfaces (VUIs) for atypical speech, enabling computers and smart devices to interpret and respond to spoken commands from speech-impaired individuals.

6.2 Deployment on a cloud computing architecture

To underscore the importance of cloud computing paradigm for our CapisciAMe project, this Section investigates the realization of a serverless architecture for atypical speech recognition in Italian.

Nowadays, the adoption of the serverless architectural pattern empowers developers to build scalable, cost-effective, and event-driven applications while reducing complexity. It offers the key benefits:

- Reduced cost. Serverless eliminates the need to provision, manage, and maintain servers. End users only pay for the resources your code uses, leading to cost savings, especially for applications with fluctuating traffic.
- Increased developer productivity. Developers can focus on writing code and building features instead of getting bogged down in server management tasks. This can significantly speed up development cycles and time to market.
- Improved scalability. Serverless architectures can automatically scale up or down based on demand. This ensures your application can handle spikes in traffic without performance degradation.
- Faster deployments. With serverless, deployments are often much faster than with traditional architectures. There's no server infrastructure to manage, so you can quickly push new code and features to production.
- Reduced latency. Serverless functions can be deployed in geographically distributed locations, bringing your code closer to users. This can significantly reduce latency and improve the responsiveness of applications.

In the depicted context, we propose to integrate our speech recognition model into a voice communicator app designed to convert the impaired speech into a more understandable human voice. This software is implemented as a web application that exploits a combination of Amazon Web Services (AWS) to pursue our goals. In particular, we use the following services over the cloud:

- Amazon Amplify: it is a development platform that simplifies building and deploying scalable and secure web applications. Amplify hosts the front-end code of our application.

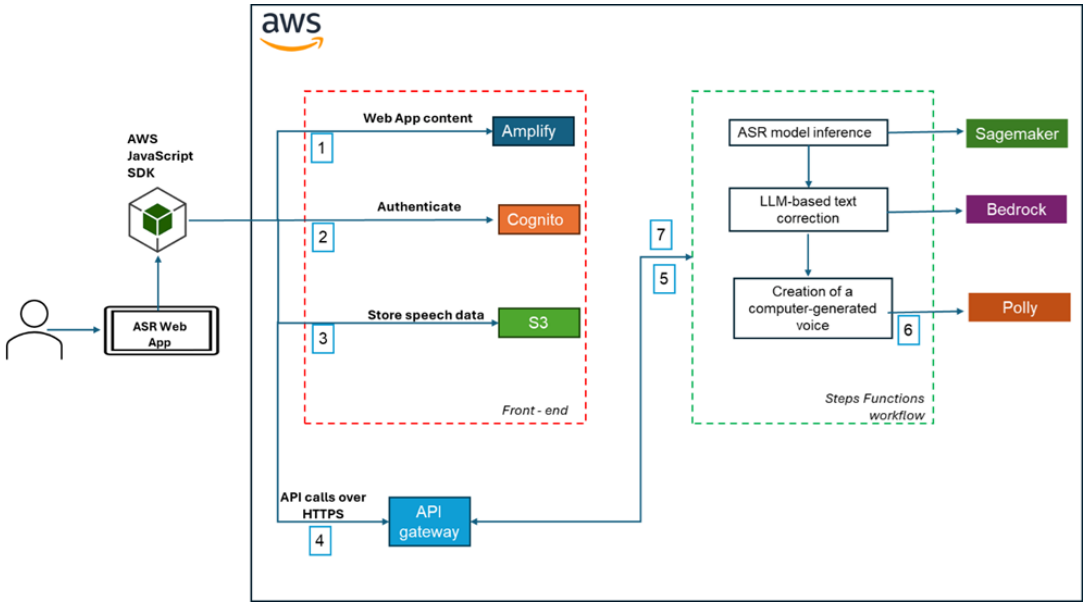


Figure 6.2: Block diagram of the voice communicator app based on AWS services

- Amazon Cognito: it is an identity platform for web and mobile apps, it serves as a user directory, an authentication server, and an authorization service.
- Amazon S3 (Simple Storage Service): it is an object storage service that offers industry-leading scalability, data availability, security, and performance. We use S3 to store the speech recordings coming from our web application.
- Amazon SageMaker: it is a fully managed service that accelerates the development, training, and deployment of machine learning (ML) models. SageMaker is the key service enabling us to execute the inference based on our fine-tuned OpenAI Whisper ASR model (see Chapter 5 for further details).
- Amazon Bedrock: it is a fully managed service that simplifies building and scaling generative AI applications using foundation models from leading AI companies, such as Meta.
- Amazon Polly: it is a cloud service that converts text into natural-sounding speech across multiple languages and voices.

Figure 6.2 describes the architecture of our application and the relationship between the considered AWS services. Here, seven key steps can be appreciated, as depicted below:

1. The user with speech impairments utilizes a web application that is compatible with computers and smart devices. The content (e.g., HTML pages, JavaScript codes) is

hosted on AWS Amplify and it employs the AWS Javascript SDK to interact with the cloud.

2. The authentication process is managed by an Amazon Cognito identity pool.

3. The speech data (e.g., WAV files) are stored in an Amazon S3 bucket.

4. The web application invokes the backend AI services by sending the Amazon S3 object key in the payload to Amazon API Gateway.

5. The API Gateway instantiates a dedicated AWS Step Functions workflow, which includes the following steps:

   - Amazon Sagemaker performs the inference on the submitted audio track, and gets the raw transcription.
   - Amazon Bedrock employs a large language model (e.g., Mistral8x7B) to revise the ASR transcription, by correcting grammatical errors and misspelled terms.
   - Amazon Polly converts the textual output (coming from the previous step) to a computer-generated voice.

6. The AWS Step Functions workflow creates an audio file as output and stores it in Amazon S3 in MP3 format.

7. The disordered speech transcription, in conjunction with a pre-signed URL with the location of the audio file stored in a dedicated Amazon S3 bucket, is sent back to the user's web app through API Gateway. In this way the user's device can play the audio file using the pre-signed URL.

Therefore, through our work, we have designed a smart VIVOCA (Voice Input Voice Output Communication Aid) that leverages CapisciAMe's recognition capabilities to support interpersonal communication for individuals with dysarthria and other speech disorders.

6.3 Smart assistance scenarios

By leveraging the synergy between Artificial Intelligence (AI) and the Internet of Things (IoT), we propose the creation of a smart assistance scenario for individuals with speech disabilities. In our vision, a smart assistance device is a connected device that interprets disordered voice commands and responds accordingly. The device utilizes its on-edge resources, along with our cloud-based transcription services, to process input speech commands and

generate external events, such as alarms or intelligent interactions with a connected home. In particular, our research study published in:

- Mulfari, D., Carnevale, L., & Villari, M. (2023). Toward a lightweight ASR solution for atypical speech on the edge. Future Generation Computer Systems, 149, 455-463. DOI: 10.1016/j.future.2023.08.002

describes the implementation of a smart assistance device powered by Raspberry Pi single-board computers, integrating suitable audio hardware (microphone and on-board speakers). Figure 6.3 illustrates the proposed prototypes. This low-cost, portable solution is mainly intended to operate next to the user with a disability, such as on a wheelchair or beside a bed, in an always-active mode [67]. Multiple contexts may benefit from the proposed approach. One notable example is achieving alternative interaction with virtual home assistants (VHAs) like Amazon Alexa. The key idea is to associate the recognition of atypical words with commands that the VHA can execute via the connected smart speaker. More specifically, with the help of Alexa-remote-control component [1], we have realized this integration. This software enables the creation of tailored shell scripts to send query commands to a selected Alexa smart speaker, so this device interprets the received command as if it were spoken by a real person.

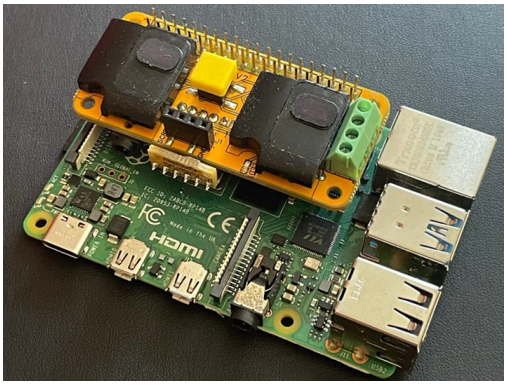
In the field of interpersonal communication, similar applications can be found. Recognizing a single atypical voice command can trigger a customized speech message, spoken aloud by a smart speaker using a computer-generated voice. This functionality can assist individuals with disabilities in clearly expressing their personal needs and enhancing interactions with caregivers and assistants, among other uses. Within these contexts, it is also important that the smart assistance devices integrate on-edge processing capabilities to perform all the speech recognition processes locally without accessing remote resources over the Internet. For these reasons, experimental evaluations were conducted with some OpenAI Whisper configurations deployed at the edge, and this study led to the results published in [109].

A working prototype of our smart assistance device is presented in this video [2].

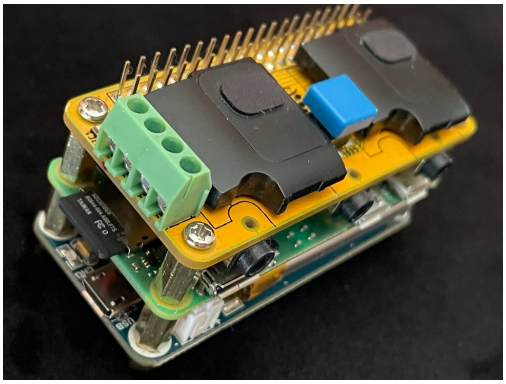
6.4 Conversational Web scenarios

Nowadays, the Conversational Web represents a significant evolution in how users interact with the internet, blending natural language processing and artificial intelligence to create

[1] https://github.com/thorsten-gehrig/alexa-remote-control
[2] https://youtu.be/S0wohiAVn88



(a) On a Raspberry Pi 4 board



(b) On a Raspberry Pi Zero board

Figure 6.3: Prototypes of smart assistance devices

more intuitive and human-like interactions. This paradigm shift aims to make web interactions as seamless and natural as having a conversation with another person. Therefore, automatic speech recognition technology plays a paramount role in empowering such hands-free interactions with machines. This is also important in the field of digital assistive technologies for people with disabilities. Following this direction, the scientific contribution in [122] proposes a novel paradigm for conversational web browsing, named ConWeb, enabling users to browse the web through conversation. As an alternative to operating graphical UIs using keyboards, mice, or screen readers, users can express their browsing goals and access websites through dialog-based interactions with a conversational agent (CA). The same research shows the effectiveness of this conversational paradigm for people with visual impairments.

Thanks to the utilization of our CapisciAMe speech recognition over the cloud, the original ConWeb paradigm and its patterns [123] have been extended to include people living with speech disorders, who can take advantage of the coordination of visual and conversational channels on the web through a multi-experience paradigm. In particular, we propose the integration of the ConWeb solution with an ASR service for dysarthric speech capable of spotting a small number of keywords within disordered utterances. This is not a limitation for our work, as it allows us to make effective use of a few impaired voice commands to explore websites proficiently.

To understand how to extend the original ConWeb paradigm to the needs of people living with dysarthria and atypical speech, we conducted a series of formative studies involving five participants with this condition, reached out through the association "TecnologicamenteInsuperabili" (under the authorization of the research-ethics committee of the Politecnico di Milano (Opinion no.11/2021)).

Our analysis identified the main challenges and related design patterns based on the coordination of a conversational paradigm adapted to the needs arising from atypical speech and a visual-augmentation paradigm highlighting labels to be pronounced to activate the "reading" of page elements. To validate the new patterns, we developed a prototype that extended the logic for conversational web browsing already implemented in ConWeb. According to this logic, the conversation is managed through a page model [123] that indexes the content segments and navigational structures available on a given web page and augments them with annotations that help build the dialog system supporting web browsing.

We also defined new customizable modules to make the ConWeb framework evolve and become flexible and adaptable to the possible variations of access paradigms that now also cover visual access. Users can now choose which combination of paradigms is most adequate, ranging from a conversation-only interaction based on natural language to the full coordination of a limited-vocabulary conversation with augmented visual browsing. This extension

involved the introduction of policies that can be configured on the client side by users, affecting how the requested pages are interpreted and rendered by the ConWeb server. We also developed a Chrome web browser extension that allows users to configure the browsing modality and manage the web page augmentation aimed at supporting the execution of the conversational agent and the visual augmentation [108].

ConWeb represents the first example of a third-party application exploiting the CapisciAMe services to improve the inclusivity and accessibility of an existing application. Further details on this research study can be found in:

- Pucci, E., Piro, L., Possaghi, I., Mulfari, D., & Matera, M. (2024). Co-designing the integration of voice-based conversational AI and web augmentation to amplify web inclusivity. Scientific Reports, 14(1), 16162.
  DOI: 10.1038/s41598-024-66725-3

6.5 Interaction with personal computers

Interaction with computers is crucial for individuals with disabilities, as it significantly enhances their independence, communication, and overall quality of life. In particular, neuromotor disabilities, such as cerebral palsy or spinal cord injuries, often impair a person's ability to perform everyday tasks using traditional input devices like keyboards, mice and touch screens. Assistive technologies and alternative input methods, such as eye-tracking, provide essential solutions.

In this context, automatic speech recognition techniques can play a crucial role as valuable alternatives to traditional input devices. We are developing a prototype Voice User Interface (VUI) for Windows PCs based on the CapisciAMe speech recognition engine. The core idea of this solution is to replace common keyboard actions with spoken commands recognized by our ASR component, which can also be optimized to run locally without requiring internet access.

Our ongoing collaboration with the ASPHI Onlus foundation in Bologna demonstrates that this methodology can enhance vocal interaction with a wide range of software that relies on predefined keystrokes to function. One domain empowered by our VUI is accessible gaming, where it can control video games that require a limited number of input commands, typically made with a computer's keyboard and gamepad. To support such real-time applications, it is paramount to optimize the impaired speech recognition process running on the local machine. An important ongoing effort focuses on optimizing ASR model inference, particularly in terms of response time.

The reported experience not only contributes to the field of assistive technology but also underscores the importance of inclusive design in creating accessible digital environments for all. A working prototype of the proposed VUI for Windows desktop environment is presented in the following video [3].

6.6 Summary

This chapter explored the implementation of speech transcription services based on our ASR engine, available over the cloud. This is a key component of the CapisciAMe ecosystem, providing software developers with an accessible platform to embed speech recognition features into custom applications. This approach not only broadens the reach of our ASR solutions but also opens avenues for supporting diverse studies and applications across various fields, such as human-computer interaction and smart assistance for people with speech disabilities.

[3] https://youtu.be/w93RAX_F9D0

Chapter 7

Conclusion and future work

This chapter summarizes the key contributions of this dissertation and proposes some possible future developments for our CapisciAMe ecosystem.

7.1 Conclusions

The research presented in this thesis lies within the domain of automatic speech recognition (ASR) and focuses on leveraging deep learning-driven techniques to support ASR services for individuals with speech impairments, such as dysarthria. Dysarthria is a prevalent neurological speech disorder that affects approximately 22 million individuals in Europe, accounting for 5% of the population. Dysarthria arises from disturbances in muscular control over the speech mechanism due to damage to the central or peripheral nervous system. This condition results in atypical speech, characterized by articulatory difficulties during oral communication caused by paralysis, weakness, or incoordination of the speech musculature. Consequently, it leads to abnormalities in the strength, speed, range, steadiness, tone, and accuracy of movements essential for speech production [1].

The aforementioned speech disorders are deeply linked with motor impairments, especially acute, as observed in cases of cerebral palsy and neurodegenerative diseases like amyotrophic lateral sclerosis, spinal muscular atrophy, multiple sclerosis, and Parkinson's disease. These conditions lead to significant motor disabilities, in conjunction with irregular speech patterns and reduced speech clarity, creating challenges for effective interpersonal communication. In particular, the impact of dysarthria on speech is particularly notable in phoneme articulation, as seen among individuals with severe disabilities. Pronounced phonemes often become indistinct, causing inaccuracies in pitch and disruptions in the articulation of vowels and consonants. These variations obscure the distinctive acoustic features necessary for ASR

systems to differentiate phonemes.

Furthermore, the diversity of disabilities among individuals with dysarthria results in speech variations that far exceed those observed in typical speech. Standard ASR systems struggle to model these variations effectively, resulting in difficulties mapping impaired speech utterances to the correct sequence of words. Addressing the needs of dysarthric individuals requires overcoming challenges related to inaccurate phonation, tempo irregularities, and inconsistencies in formant transitions. Indeed, multiple challenges face contemporary ASR solutions when dealing with impaired voices, including the alternation and inaccuracy of dysarthric phonemes, the scarcity of dysarthric speech data, and the phoneme labeling imprecision due to the neurological disorder [3].

For this reason, automatic disordered speech recognition has garnered considerable attention from both industry and research communities. While global efforts and significant research projects are focused on English and American languages, our PhD research activity has centered on realizing AI-based automatic speech recognition solutions to aid individuals with disordered speech who speak Italian. Our project is centered on the real needs of individuals experiencing severe motor disabilities and dysarthria.

In order to address the various challenges in our research, we propose a technological approach following assistive technology principles. We first simplify the challenge by excluding issues in continuous and connected speech recognition in the presence of disorders. Following this direction, it is important to have an ASR system capable of recognizing a limited number of isolated words and short meaningful sentences pronounced by selected speakers with disordered speech. More specifically, we propose the design and initial implementation of a digital ecosystem supporting ASR services in Italian, named CapisciAMe. Its complex architecture is composed of three interconnected pillars, which are crucial aspects of our research:

- Disordered speech collection. Given the scarcity of dysarthric corpora, especially in Italian, a key part of our work involves acquiring voice samples from individuals with speech disorders, as discussed in Chapter 4. Over the years, this ongoing effort has culminated in the creation of the first Italian atypical corpus for AI-based research, along with a closed-set vocabulary. To achieve this, we introduced novel software assistive technology solutions, such as the CapisciAMe mobile app, to facilitate speech acquisition from end users and procedures to enhance impaired speech signals. This has led to the creation and expansion of the CapisciAMe database, which now represents the richest and most complete corpus of voice samples from people with speech disorders in Italian, to the best of our knowledge. It encompasses more than 90 hours of labeled atypical speech, contributed by 250 anonymized speech-impaired users globally.

- Deep learning architectures. We explore models specifically designed for pure keyword spotting tasks and sequence-to-sequence structures that exploit sophisticated pre-existing DL frameworks, known as encoder-decoder models, which are currently state-of-the-art (SOTA) architectures, namely Wav2Vec2 and Whisper. These models are pre-trained on extensive datasets of multilingual standard speech and are fine-tuned on the entire CapisciAMe corpus. This is of paramount importance for our work, as it allows us to recognize single voice commands and short sentences (composed of a combination of isolated words) uttered by speakers with atypical speech and dysarthria. In addition, our experimental evaluations, described in detail in Chapter 5, have confirmed the effectiveness of our design choices in terms of recognition accuracy, with a word error rate of 3.5%. This represents a very competitive result against the automatic disordered speech recognition solutions available in the recent literature [30].

- ASR services and application prototypes. We propose integrating our ASR system into Voice User Interfaces (VUIs). Deploying our speech recognizer as a cloud-based service is crucial, as it facilitates real-world applications that harness the ASR system's capabilities. Therefore, this on-demand speech-to-text conversion plays a key role in our technological ecosystem since it provides software developers with a convenient way to integrate VUI into their custom applications. The availability of these ASR remote resources can significantly support other studies exploiting automatic disordered speech recognition in diverse fields, such as e-health. This represents a key finding of our work, enabling progress in the field of speech recognition for atypical voices.

Specifically, our work intends to answer the following research questions:

RQ1: How should a mobile voice collection system be designed?

Our project innovatively uses mobile devices with the CapisciAMe app to collect speech samples from individuals with dysarthria and other speech disorders. The app features a simple, intuitive interface, making it accessible, especially for users with motor impairments. Users log in with a nickname, optionally sharing personal information, and all data is securely stored in a private cloud repository. Our solution utilizes common smartphones, integrating necessary components for recording voice samples and supporting external microphones for high-quality recordings. It allows users to collect utterances anytime and anywhere, leveraging the portability of mobile devices. The app is designed to encourage users to share their speech samples by making the process straightforward and rewarding. Users can test the speech recognition capabilities of the ASR engine, validating their contributions. CapisciAMe is continuously improved based on user feedback and performance metrics, adapting

to new technologies and user needs over time.

RQ2: What are the benefits of using state-of-the-art ASR models in automatic disordered speech recognition?

In this thesis, our more performant speech recognition system is based on sequence-to-sequence structures that exploit sophisticated pre-existing DL frameworks, known as encoder-decoder models, namely Wav2Vec2 and Whisper. This technology processes and translates spoken language into text by mapping sequences of speech frames directly to sequences of characters. The models are pre-trained on extensive datasets of multilingual standard speech and fine-tuned on the entire CapisciAMe corpus. This is crucial for our work, as it allows us to recognize single voice commands and short sentences (composed of isolated words) uttered by speakers with atypical speech. Experimental evaluations demonstrated the effectiveness of the proposed approaches in terms of word error rate, as detailed in Chapter 5.

RQ3: What are the potential applications of voice user interfaces for atypical speech?

As discussed in Chapter 6, the designed speech recognition engine can empower a wide range of applications for people with speech disorders. Examples include interpersonal communication, human-computer interaction, smart home automation, e-health, and tele-rehabilitation scenarios.

7.2 Future works

The previous section summarized the current status of our research in the field of automatic speech recognition in the presence of dysarthria and other speech disorders. However, some challenges still need to be addressed to improve the overall performance of the proposed CapisciAMe digital ecosystem. Future directions for this work are presented below and concern all parts of our current ASR architecture.

7.2.1 Speech database extension

At the time of writing, CapisciAMe is an ongoing project, with its core being our private corpus of disordered speech in Italian. The availability of this data is of paramount importance for our AI research, as it allows us to train ASR speech models. To enhance the robustness and generalizability of our ASR models, it is crucial to increase the size and diversity of the CapisciAMe database. Collecting more voice samples from individuals with speech disorders caused by various conditions and states is essential. Diversity in data ensures that our

models can effectively handle the wide range of speech variations present in different types of speech impairments. This diversity helps improve the accuracy and reliability of ASR systems, making them more inclusive and effective for all users, regardless of the specific nature of their speech disorder.

In this context, we believe that collaborating with hospitals, rehabilitation centers, and associations of patients is crucial for conducting a well-structured speech collection. Hospitals and patient associations can provide access to a diverse group of individuals with various speech disorders. This diversity is essential for creating a comprehensive and representative speech corpus, which in turn enhances the robustness and generalizability of our ASR models. Collaborating with established medical institutions ensures that the speech collection process adheres to ethical standards. Hospitals and patient associations can help facilitate informed consent from participants, ensuring that they understand the purpose of the research and their rights.

Healthcare professionals can offer valuable insights into the specific characteristics of different speech disorders. This expertise can guide the design of the speech collection process, ensuring that it captures relevant and meaningful data for ASR model training. Patients may feel more comfortable participating in research facilitated by trusted medical institutions and associations. This trust can lead to higher participation rates and more natural speech samples, which are critical for developing effective ASR systems. Therefore, this collaboration is key to advancing our research and developing ASR solutions that truly meet the needs of individuals with speech disorders.

7.2.2 Enhancements in speech recognition engine

As discussed in Chapter 5, the utilization of state-of-the-art ASR architectures, which are based on sequence-to-sequence pre-trained models fine-tuned on our impaired speech corpus, has allowed us to recognize both isolated words and short sentences pronounced by individuals with speech disorders. In particular, experimental evaluation has shown the remarkable performance of OpenAI Whisper in recognition accuracy, achieving an overall WER of 3.5% on a custom testing dataset. These results are significant and represent a small but important step toward continuous speech recognition. To support this process, we believe that correcting any malformed strings generated by our Whisper fine-tuned model is crucial.

Following this direction, we believe that our comprehensive application can benefit from large language models (LLMs). LLMs can help improve the robustness of our ASR system by correcting malformed strings and handling variations in speech. Their ability to understand context and predict likely word sequences allows them to correct errors that may

arise from speech impairments, leading to more accurate transcriptions. As we collect more data and refine our models, LLMs can be continuously updated and fine-tuned to improve their performance. This adaptability ensures that our ASR system remains state-of-the-art and can handle new and diverse speech patterns as they are encountered. Furthermore, LLMs can be integrated with other assistive technologies, such as text-to-speech systems and communication aids, to create a comprehensive support system for individuals with speech disorders. This integration may enhance the overall user experience by providing more effective communication tools.

At the same time, as speech recognition technology advances, in future works we intend to explore the utilization of state-of-the-art technology on disordered speech. We also plan to investigate the same CapisciAMe methodology in different languages, such as English.

7.2.3 Applications based on the CapisciAMe speech recognition engine

As discussed in the previous chapter, both the integration of the CapisciAMe ASR capabilities into real-world scenarios and the development of novel applications leveraging our speech recognition services play a fundamental role in our project. To support this aspect, we have deployed our ASR solution as an on-demand service, enabling third parties to implement voice user interfaces for atypical speech into custom applications. We believe that the practical use of these solutions can incentivize speech collection from end users, helping us to overcome the challenges in expanding our private corpus of disordered speech in Italian.

Following this direction, in future works, we plan to consolidate our current prototypes with the aim of turning them into concrete aids that support the everyday activities of people with speech disabilities, particularly in the field of human-computer interaction. An interesting area of future research and application lies in developing real-time ASR capabilities to enable immediate transcription and feedback for users with speech impairments. This could be particularly useful in live communication settings, such as video calls or in-person conversations, and could be exploited as a VIVOCA (Voice Input Voice Output Communication Aid).

7.2.4 Concluding remarks

This thesis explores AI-driven techniques to support ASR services for individuals with speech disorders, particularly dysarthria, a neurological disorder affecting speech motor mechanisms and leading to poor intelligibility. We propose the design and initial implementation of the CapisciAMe digital ecosystem for automatic disordered speech recognition in Italian, focusing

on precise disordered speech command recognition using speaker-dependent approaches and advanced deep learning solutions, including Transformer-based state-of-the-art ASR models, enhanced by fine-tuning techniques. This study makes significant contributions by:

- Applying cutting-edge ASR technology to Italian impaired speech.
- Developing and deploying customized ASR models for cloud-based and edge use.
- Creating a valuable resource for dysarthric speech research through the CapisciAMe private database, which now represents the richest and most complete corpus of voice samples from people with speech disorders in Italian.

In addition, our research demonstrates the feasibility of creating ASR systems for disordered speech by addressing challenges in data scarcity, model optimization, and application accessibility. The CapisciAMe ecosystem achieves significant improvements in recognizing impaired speech in Italian, laying the groundwork for further development of inclusive communication technologies that enhance the independence and social participation of individuals with speech impairments.

Bibliography

[1] Joseph R Duffy. *Motor speech disorders e-book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2019.

[2] Luigi De Russis and Fulvio Corno. On the impact of dysarthric speech on contemporary asr cloud platforms. *Journal of Reliable Intelligent Environments*, 5(3):163–172, 2019.

[3] Seyed Reza Shahamiri. Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:852–861, 2021.

[4] Aisha Jaddoh, Fernando Loizides, and Omer Rana. Interaction between people with dysarthria and speech recognition systems: A review. *Assistive Technology*, pages 1–9, 2022.

[5] Feifei Xiong, Jon Barker, and Heidi Christensen. Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5836–5840. IEEE, 2019.

[6] Fabio Ballati, Fulvio Corno, and Luigi De Russis. Assessing virtual assistant capabilities with italian dysarthric speech. pages 93–101. Association for Computing Machinery, 2018.

[7] Davide Mulfari, Gabriele Meoni, and Luca Fanucci. Machine learning in assistive technology: a solution for people with dysarthria. In *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*, pages 308–309, 2018.

[8] Rabbia Mahum, Ahmed M El-Sherbeeny, Khaled Alkhaledi, and Haseeb Hassan. Trandsr: A hybrid model for dysarthric speech recognition using transformer encoder and ensemble learning. *Applied Acoustics*, 222:110019, 2024.

[9] Aisha Jaddoh, Fernando Loizides, Omer Rana, and Yasir Ahmed Syed. Interacting with smart virtual assistants for individuals with dysarthria: A comparative study on usability and user preferences. *Applied Sciences*, 14(4), 2024.

[10] Mark Hasegawa-Johnson, Xiuwen Zheng, Heejin Kim, Clarion Mendes, Meg Dickinson, Erik Hege, Chris Zwilling, Marie Moore Channell, Laura Mattie, Heather Hodges, et al. Community-supported shared infrastructure in support of speech accessibility. *Journal of Speech, Language, and Hearing Research*, pages 1–14, 2024.

[11] Dave L Edyburn. Rethinking assistive technology. *Special Education Technology Practice*, 5(4):16–23, 2004.

[12] Zhaopeng Qian, Kejing Xiao, and Chongchong Yu. A survey of technologies for automatic dysarthric speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1):48, 2023.

[13] Carole Roth. *Dysarthria*, pages 905–908. Springer New York, New York, NY, 2011.

[14] Pam Enderby. Disorders of communication: dysarthria. *Handbook of clinical neurology*, 110:273–281, 2013.

[15] Cristina Mei, Sheena Reilly, Dinah Reddihough, Fiona Mensah, and Angela Morgan. Motor speech impairment, activity, and participation in children with cerebral palsy. *International Journal of Speech-Language Pathology*, 16(4):427–435, 2014.

[16] Claire Mitchell, Matthew Gittins, Sarah Tyson, Andy Vail, Paul Conroy, Lizz Paley, and Audrey Bowen. Prevalence of aphasia and dysarthria among inpatient stroke survivors: Describing the population, therapy provision and outcomes on discharge. *Aphasiology*, 35(7):950–960, 2021.

[17] Jimin Lee, Aarthi Madhavan, Elizabeth Krajewski, and Sydney Lingenfelter. Assessment of dysarthria and dysphagia in patients with amyotrophic lateral sclerosis: Review of the current evidence. *Muscle & Nerve*, 2021.

[18] Laureano Moro-Velazquez, Jorge A Gomez-Garcia, Julian D Arias-Londoño, Najim Dehak, and Juan I Godino-Llorente. Advances in parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects. *Biomedical Signal Processing and Control*, 66:102418, 2021.

[19] Stephanie H Felgoise, Vincenzo Zaccheo, Jason Duff, and Zachary Simmons. Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 17(3-4):179–183, 2016.

[20] Alan Wrench and Korin Richmond. Continuous speech recognition using articulatory data. In *Sixth International Conference on Spoken Language Processing (ICSLP 2000)*, pages 145–148. International Speech Communication Association, 2000.

[21] Aisha Jaddoh, Fernando Loizides, Jimin Lee, and Omer Rana. An interaction framework for designing systems for virtual home assistants and people with dysarthria. *Universal Access in the Information Society*, pages 1–13, 2023.

[22] Victoria Young and Alex Mihailidis. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2):99–112, 2010.

[23] Davide Mulfari, Gabriele Meoni, Marco Marini, and Luca Fanucci. Machine learning assistive application for users with speech disorders. *Applied Soft Computing*, 103:107147, 2021.

[24] Zhaopeng Qian and Kejing Xiao. A survey of automatic speech recognition for dysarthric speech. *Electronics*, 12(20), 2023.

[25] Komal Bharti and Pradip K. Das. A survey on asr systems for dysarthric speech. In *2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST)*, pages 1–6, 2022.

[26] Mark S. Hawley, Pam Enderby, Phil Green, Stuart Cunningham, and Rebecca Palmer. Development of a voice-input voice-output communication aid (vivoca) for people with severe dysarthria. In Klaus Miesenberger, Joachim Klaus, Wolfgang L. Zagler, and Arthur I. Karshmer, editors, *Computers Helping People with Special Needs*, pages 882–885, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[27] Stuart Cunningham, Phil Green, Heidi Christensen, José Atria, Andre Coy, Massimiliano Malavasi, Lorenzo Desideri, and Frank Rudzicz. Cloud-based speech technology for assistive technology applications (cloudcast). volume 242, 01 2017.

[28] Massimiliano Malavasi, Enrico Turri, Jose Atria, Heidi Christensen, Ricard Marxer, Lorenzo Desideri, Andre Coy, Fabio Tamburini, and Phil Green. An innovative speech-

based user interface for smarthomes and iot solutions to help people with speech and motor disabilities. *Studies in health technology and informatics*, 242:306, 09 2017.

[29] Massimiliano Donati, Alessio Bechini, Clelia D'Anna, Bruno Fattori, Marco Marini, Martina Olivelli, Susanna Pelagatti, Giulia Ricci, Erika Schirinzi, Gabriele Siciliano, Mirko Tavosanis, Francesca Torri, Nicola Vanello, and Luca Fanucci. A clinical tool for prognosis and speech rehabilitation in dysarthric patients: The desire project. In Riccardo Berta and Alessandro De Gloria, editors, *Applications in Electronics Pervading Industry, Environment and Society*, pages 380–385, Cham, 2023. Springer Nature Switzerland.

[30] Seyed Reza Shahamiri, Vanshika Lal, and Dhvani Shah. Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3407–3416, 2023.

[31] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S. Huang, Kenneth Watkin, and Simone Frame. Dysarthric speech database for universal access research. In *Proc. Interspeech 2008*, pages 1741–1744, 2008.

[32] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46:523–541, 2012.

[33] Ray D Kent, Gary Weismer, Jane F Kent, and John C Rosenbek. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54(4):482–499, 1989.

[34] Kathryn M Yorkston, David R Beukelman, F Minifie, and Shimon Sapir. Assessment of stress patterning. *The dysarthria: Physiology, acoustics, perception, management*, pages 131–162, 1984.

[35] Louis CW Pols, Xue Wang, and Louis FM ten Bosch. Modelling of phone duration (using the timit database) and its potential benefit for asr. *Speech Communication*, 19(2):161–176, 1996.

[36] Christian Kroos. Measurement accuracy in 3d electromagnetic articulography (carstens ag500). In *Proceedings of the 8th international seminar on speech production*, pages 61–64, 2008.

[37] Nada Gohider and Otman A Basir. Recent advancements in automatic disordered speech recognition: A survey paper. *Natural Language Processing Journal*, page 100110, 2024.

[38] Bob MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn Ladewig, Jimmy Tobin, Michael Brenner, Philip Q Nelson, et al. Disordered speech data collection: Lessons learned at 1 million utterances from project euphonia. 2021.

[39] Mauro Nicolao, Heidi Christensen, Stuart Cunningham, Phil Green, and Thomas Hain. A framework for collecting realistic recordings of dysarthric speech-the homeservice corpus. In *Proceedings of LREC 2016*. European Language Resources Association, 2016.

[40] Xavier Menendez-Pidal, James B Polikoff, Shirley M Peters, Jennie E Leonzio, and H Timothy Bunnell. The nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1962–1965. IEEE, 1996.

[41] Wendell Johnson, Frederic L Darley, and Duane C Spriesterbach. Diagnostic methods in speech pathology. 1963.

[42] JR Deller Jr, MS Liu, LJ Ferrier, and P Robichaud. The whitaker database of dysarthric (cerebral palsy) speech. *The Journal of the Acoustical Society of America*, 93(6):3516–3518, 1993.

[43] Mengyi Sun, Ming Gao, Xinchen Kang, Shiru Wang, Jun Du, Dengfeng Yao, and Su-Jing Wang. Cdsd: Chinese dysarthria speech database. *arXiv preprint arXiv:2310.15930*, 2023.

[44] Emre Yilmaz, MS Ganzeboom, LJ Beijer, Catia Cucchiarini, and Helmer Strik. A dutch dysarthric speech database for individualized speech therapy research. 2016.

[45] Cécile Fougeron, Lise Crevier-Buchman, Corinne Fredouille, Alain Ghio, Christine Meunier, Claude Chevrie-Muller, Nicolas Audibert, Jean-François Bonastre, Antonia Colazo-Simon, Céline Delooze, et al. Developing an acoustic-phonetic characterization of dysarthric speech in french. In *7th International Conference on Language Resources, Technologies and Evaluation (LREC)*, volume 1, pages 2831–2838. Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Bente Maegaard ..., 2010.

[46] Gwen Van Nuffelen, Marc De Bodt, Catherine Middag, and Jean-Pierre Martens. Dutch corpus of pathological and normal speech (copas). *Antwerp University Hospital and Ghent University, Tech. Rep*, 2009.

[47] Pavel Grill and Jana Tučková. Speech databases of typical children and children with sli. *PloS one*, 11(3):e0150365, 2016.

[48] Christine Meunier, Cécile Fougeron, Corinne Fredouille, Brigitte Bigi, Lise Crevier-Buchman, Elisabeth Delais-Roussarie, Laurianne Georgeton, Alain Ghio, Imed Laaridh, Thierry Legou, et al. The typaloc corpus: A collection of various dysarthric speech recordings in read and spontaneous styles. In *Language Resources and Evaluation Conference (LREC)*, pages p–4658, 2016.

[49] Eun Jung Yeo, Sunhee Kim, and Minhwa Chung. Automatic severity classification of korean dysarthric speech using phoneme-level pronunciation features. In *Interspeech*, pages 4838–4842, 2021.

[50] Ka Ho Wong, Yu Ting Yeung, Edwin HY Chan, Patrick CM Wong, Gina-Anne Levow, and Helen Meng. Development of a cantonese dysarthric speech corpus. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[51] Marco Marini, Mauro Vigano, Massimo Corbo, Marina Zettin, Gloria Simoncini, Bruno Fattori, Clelia D'Anna, Massimiliano Donati, and Luca Fanucci. Idea: An italian dysarthric speech database. *2021 IEEE Spoken Language Technology Workshop, SLT 2021 - Proceedings*, pages 1086–1093, 1 2021.

[52] Rosanna Turrisi, Arianna Braccia, Marco Emanuele, Simone Giulietti, Maura Pugliatti, Mariachiara Sensi, Luciano Fadiga, and Leonardo Badino. Easycall corpus: a dysarthric speech dataset. 2021.

[53] Melanie Fried-Oken. Voice recognition device as a computer interface for motor and speech impaired people. *Archives of physical medicine and rehabilitation*, 66(10):678–681, 1985.

[54] Linda Ferrier, Howard Shane, Holly Ballard, Tyler Carpenter, and Anne Benoit. Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication*, 11(3):165–175, 1995.

[55] Ming Tu, Alan Wisler, Visar Berisha, and Julie M Liss. The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance. *The Journal of the Acoustical Society of America*, 140(5):EL416–EL422, 2016.

[56] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[57] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[58] JR Deller Jr, D Hsu, and Linda J Ferrier. On the use of hidden markov modelling for recognition of dysarthric speech. *Computer Methods and Programs in Biomedicine*, 35(2):125–139, 1991.

[59] N Rajeswari and S Chandrakala. Generative model-driven feature learning for dysarthric speech recognition. *Biocybernetics and Biomedical Engineering*, 36(4):553–561, 2016.

[60] Harsh Vardhan Sharma and Mark Hasegawa-Johnson. State-transition interpolation and map adaptation for hmm-based dysarthric speech recognition. In *Proceedings of the NAACL HLT 2010 workshop on speech and language processing for assistive technologies*, pages 72–79, 2010.

[61] Seyed Reza Shahamiri and Siti Salwah Binti Salim. Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of mfcc parameters and studying a speaker-independent approach. *Advanced Engineering Informatics*, 28(1):102–110, 2014.

[62] Seyed Reza Shahamiri. Neural network-based multi-view enhanced multi-learner active learning: theory and experiments. *Journal of Experimental & Theoretical Artificial Intelligence*, 34(6):989–1009, 2022.

[63] S Selva Nidhyananthan, R Shantha Selva kumari, and V Shenbagalakshmi. Assessment of dysarthric speech using elman back propagation network (recurrent network) for speech recognition. *International Journal of Speech Technology*, 19:577–583, 2016.

[64] Brahim Fares Zaidi, Sid Ahmed Selouani, Malika Boudraa, and Mohammed Sidi Yakoub. Deep neural network architectures for dysarthric speech analysis and recognition. *Neural Computing and Applications*, 33(15):9089–9108, 2021.

[65] Prasad D Polur and Gerald E Miller. Investigation of an hmm/ann hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals. *Medical engineering & physics*, 28(8):741–748, 2006.

[66] Davide Mulfari, Lorenzo Carnevale, and Massimo Villari. Toward a lightweight asr solution for atypical speech on the edge. *Future Generation Computer Systems*, 149:455–463, 2023.

[67] Davide Mulfari, Lorenzo Carnevale, Antonino Galletta, and Massimo Villari. Edge computing solutions supporting voice recognition services for speakers with dysarthria. In *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW)*, pages 231–236, 2023.

[68] Davide Mulfari, Donatella La Placa, Chiara Rovito, Antonio Celesti, and Massimo Villari. Deep learning applications in telerehabilitation speech therapy scenarios. *Computers in Biology and Medicine*, 148:105864, 2022.

[69] Davide Mulfari, Gabriele Meoni, Marco Marini, and Luca Fanucci. Towards a deep learning based asr system for users with dysarthria. In *International Conference on Computers Helping People with Special Needs*, pages 554–557. Springer, 2018.

[70] I Sutskever. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[71] Yu Zhang, William Chan, and Navdeep Jaitly. Very deep convolutional networks for end-to-end speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4845–4849. IEEE, 2017.

[72] Renyuan Liu, Jian Yang, and Mengyuan Liu. A new end-to-end long-time speech synthesis system based on tacotron2. In *Proceedings of the 2019 international symposium on signal processing systems*, pages 46–50, 2019.

[73] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE, 2016.

[74] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4945–4949. IEEE, 2016.

[75] Sreenivas Sremath Tirumala and Seyed Reza Shahamiri. A deep autoencoder approach for speaker identification. In *Proceedings of the 9th international conference on signal processing systems*, pages 175–179, 2017.

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[77] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 449–456. IEEE, 2019.

[78] Seyed Reza Shahamiri, Wan MN Wan Kadir, and Suhaimi Ibrahim. A single-network ann-based oracle to verify logical software modules. In *2010 2nd international conference on software technology and engineering*, volume 2, pages V2–272. IEEE, 2010.

[79] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5884–5888. IEEE, 2018.

[80] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

[81] Jimmy Tobin and Katrin Tomanek. Personalized automatic speech recognition trained on small disordered speech datasets. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6637–6641. IEEE, 2022.

[82] Jordan R Green, Robert L MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A Ladewig, Jimmy Tobin, Michael P Brenner, Philip C Nelson, and Katrin Tomanek. Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. pages 4778–4782, 2021.

[83] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*, 2021.

[84] Bo Li, Ruoming Pang, Yu Zhang, Tara N Sainath, Trevor Strohman, Parisa Haghani, Yun Zhu, Brian Farris, Neeraj Gaur, and Manasa Prasad. Massively multilingual asr: A lifelong learning solution. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6397–6401. IEEE, 2022.

[85] Rishabh Jain, Andrei Barcovschi, Mariam Yahayah Yiwere, Dan Bigioi, Peter Corcoran, and Horia Cucu. A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition. *IEEE Access*, 11:46938–46948, 2023.

[86] Juan Camilo Vásquez-Correa and Aitor Álvarez Muniain. Novel speech recognition systems applied to forensics within child exploitation: Wav2vec2. 0 vs. whisper. *Sensors*, 23(4):1843, 2023.

[87] Ahmed Adel Attia, Jing Liu, Wei Ai, Dorottya Demszky, and Carol Espy-Wilson. Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 74–80, 2024.

[88] Agnes Luhtaru, Rauno Jaaska, Karl Kruusamäe, and Mark Fishel. Automatic transcription for estonian children's speech. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 705–709, 2023.

[89] Andrei Barcovschi, Rishabh Jain, and Peter Corcoran. A comparative analysis between conformer-transducer, whisper, and wav2vec2 for improving the child speech recognition. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 42–47, 2023.

[90] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE, 2021.

[91] Pu Wang and Hugo Van Hamme. Benefits of pre-trained mono-and cross-lingual speech representations for spoken language understanding of dutch dysarthric speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1):15, 2023.

[92] Tatsunari Matsushima. *Dutch Dysarthric Speech Recognition: Applying Self-Supervised Learning to Overcome the Data Scarcity Issue*. PhD thesis, 2022.

[93] Murali Karthick Baskar, Tim Herzig, Diana Nguyen, Mireia Diez, Tim Polzehl, Lukáš Burget, Jan Černockỳ, et al. Speaker adaptation for wav2vec2 based dysarthric asr. *arXiv preprint arXiv:2204.00770*, 2022.

[94] Lidan Wu, Daoming Zong, Shiliang Sun, and Jing Zhao. A sequential contrastive learning framework for robust dysarthric speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7303–7307. IEEE, 2021.

[95] Lester Phillip Violeta, Wen-Chin Huang, and Tomoki Toda. Investigating self-supervised pretraining frameworks for pathological speech recognition. *arXiv preprint arXiv:2203.15431*, 2022.

[96] Abner Hernandez, Paula Andrea Pérez-Toro, Elmar Nöth, Juan Rafael Orozco-Arroyave, Andreas Maier, and Seung Hee Yang. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition. *arXiv preprint arXiv:2204.01670*, 2022.

[97] Iván G. Torre, Mónica Romero, and Aitor Álvarez. Improving aphasic speech recognition by using novel semi-supervised learning methods on aphasiabank for english and spanish. *Applied Sciences*, 11(19), 2021.

[98] Giulia Sanguedolce, Patrick A. Naylor, and Fatemeh Geranmayeh. Uncovering the potential for a weakly supervised end-to-end model in recognising speech from patient with post-stroke aphasia. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 182–190, Toronto, Canada, July 2023. Association for Computational Linguistics.

[99] Siddharth Rathod, Monil Charola, and Hemant A Patil. Transfer learning using whisper for dysarthric automatic speech recognition. In *International Conference on Speech and Computer*, pages 579–589. Springer, 2023.

[100] R Vinotha, D Hepsiba, and LD Vijay Anand. Leveraging openai whisper model to improve speech recognition for dysarthric individuals. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, pages 1–5. IEEE, 2024.

[101] Shujie Hu, Xurong Xie, Zengrui Jin, Mengzhe Geng, Yi Wang, Mingyu Cui, Jiajun Deng, Xunying Liu, and Helen Meng. Exploring self-supervised pre-trained asr models

for dysarthric and elderly speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[102] Richard Cave. How people living with amyotrophic lateral sclerosis use personalized automatic speech recognition technology to support communication. *Journal of Speech, Language, and Hearing Research*, pages 1–17, 2024.

[103] Douglas O'Shaughnessy. Speech enhancement—a review of modern methods. *IEEE Transactions on Human-Machine Systems*, 54(1):110–120, 2024.

[104] Jean-Marc Valin. A hybrid dsp/deep learning approach to real-time full-band speech enhancement. In *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*, pages 1–5. IEEE, 2018.

[105] Davide Mulfari, Giuseppe Campobello, Giovanni Gugliandolo, Antonio Celesti, Massimo Villari, and Nicola Donato. Comparison of noise reduction techniques for dysarthric speech recognition. In *2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6, 2022.

[106] Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 146–151. IEEE, 2023.

[107] Davide Mulfari, Antonio Celesti, and Massimo Villari. Exploring ai-based speaker dependent methods in dysarthric speech recognition. In *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pages 958–964, 2022.

[108] Emanuele Pucci, Ludovica Piro, Isabella Possaghi, Davide Mulfari, and Maristella Matera. Co-designing the integration of voice-based conversational ai and web augmentation to amplify web inclusivity. *Scientific Reports*, 14(1):16162, 2024.

[109] Davide Mulfari and Massimo Villari. A voice user interface on the edge for people with speech impairments. *Electronics*, 13(7), 2024.

[110] Davide Mulfari, Lorenzo Carnevale, and Massimo Villari. Sequence-to-sequence models in italian atypical speech recognition. In *2024 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6, 2024.

[111] Ahmad Almadhor, Rizwana Irfan, Jiechao Gao, Nasir Saleem, Hafiz Tayyab Rauf, and Seifedine Kadry. E2e-dasr: End-to-end deep learning-based dysarthric automatic speech recognition. *Expert Systems with Applications*, 222:119797, 2023.

[112] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[113] C. Sireesha, K. Asish, and R. Manjula. A fine-tuned transformer model for dysarthric speech with spectrograms. In *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, pages 1–6, 2024.

[114] Tara N Sainath and Carolina Parada. Convolutional neural networks for small-footprint keyword spotting. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[115] Namrata Dave. Feature extraction methods lpc, plp and mfcc in speech recognition. *International journal for advance research in engineering and technology*, 1(6):1–4, 2013.

[116] Siddique Latif, Aun Zaidi, Heriberto Cuayahuitl, Fahad Shamshad, Moazzam Shoukat, and Junaid Qadir. Transformers in speech processing: A survey. *arXiv preprint arXiv:2303.11607*, 2023.

[117] Hemant Yadav and Sunayana Sitaram. A survey of multilingual models for automatic speech recognition. *arXiv preprint arXiv:2202.12576*, 2022.

[118] Yunpeng Liu, Xukui Yang, and Dan Qu. Exploration of whisper fine-tuning strategies for low-resource asr. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29, 2024.

[119] DN Krishna, Pinyi Wang, and Bruno Bozza. Using large self-supervised models for low-resource speech recognition. In *Interspeech*, pages 2436–2440, 2021.

[120] Andrew Cameron Morris, Viktoria Maier, and Phil D Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768, 2004.

[121] Ir Theo Bougie. Iso 9999 assistive products for persons with disability—classification and terminology. *The engineering handbook of smart technology for aging, disability and independence*, pages 117–126, 2008.

[122] Marcos Baez, Claudia Maria Cutrupi, Maristella Matera, Isabella Possaghi, Emanuele Pucci, Gianluca Spadone, Cinzia Cappiello, and Antonella Pasquale. Exploring challenges for conversational web browsing with blind and visually impaired users. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.

[123] Emanuele Pucci, Isabella Possaghi, Claudia Maria Cutrupi, Marcos Baez, Cinzia Cappiello, and Maristella Matera. Defining patterns for a conversational web. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.