POLITECNICO DI TORINO

Master's Degree in Software Engineering

Master's Degree Thesis

Voice segmentation for early identification of neurodegenerative diseases

Supervisors
prof. Gabriella Olmo
dr. Federica Amato

Candidate
Luca Rinelli
ID 269097

Academic year 2021-2022

Abstract

In the context of neurodegenerative diseases, vocal analysis is an easy and inexpensive way to extract useful elements for follow-up, prognostic prediction and rehabilitation of patients. Datasets with recordings of patients and healthy controls reading prompts have been collected to allow the development and evaluation of automatic vocal analysis systems.

This work deals with the tasks of speech recognition and forced alignment, important building blocks for such systems, specifically on speech from Parkinson's Disease patients and healthy controls.

A system has been designed and developed to perform automatic speech recognition and forced alignment, based on fine-tuned state-of-the-art models, it can leverage unlabeled data, and it can be used to align spontaneous speech with no prompts. The output of the system are the words and phonemes identified in a recording and the time-alignment of each single word and phoneme; this enables the automatic segmentation of vocal data and adds other data points for subsequent steps of analysis and correlation with the clinical parameters of the patients. The performance of the system is evaluated on "normal" speech datasets and disordered speech.

Contents

List of Tables

List of Figures

Part I

Introduction

Chapter 1

Introduction

Parkinson's Disease (PD) is the second most common Neurodegenerative Disease (ND), affecting 6.1 million individuals in 2016[52] and predicted to affect more than 9 million by 2030[139].

Currently, no reliable test yet exists to spot Alzheimer's Disease (AD) or PD before symptoms appear. And, there are neither laboratory nor instrumental tests for monitoring the disease[50] that would be useful during treatment. Moreover, there is no cure for PD, but it is expected that new drugs will be able to stop its advance or at least slow it down enough to extend significantly the period in which the patient is independent[120]. In this context, an early diagnosis will be crucial. Indeed, emerging evidence suggests that voice dysfunction is the earliest sign of motor impairment in PD[108] and it can be noticed up to 10 years before diagnosis[85, 140].

A growing body of literature in the past decades[119, 125] has been focusing on speech and voice impairment of people with PD, mainly including automatic and computerized analysis. Some of these approaches analyze features related to specific phonetic groups in continuous speech, therefore they need to perform first speech segmentation. This operation is often performed manually by a human operator, reducing the scalability of the approach and increasing their cost.

Automatic speech segmentation or Forced Alignment (FA) has been widely used in sociolinguistics, phonetics, language documentation, and psycholinguistics[114]. Applying FA to the speech of PD patients however presents some difficulties. It is hard to have a patient to read clearly and precisely from a prompt. Typically, recordings contain repetitions, skipping of words or syllables, mispronunciation, insertion of superfluous phonemes. This unexpected sounds cause most FA software to output wrong alignments, often for the entire recording that is being aligned.

The goal of this work is to realize a FA system for speech in Italian, that is robust to the unexpected sounds just described, to be used as building block for systems of automatic diagnosis or monitoring of PD through recordings of continuous speech.

1.1 Objectives and hypothesis

The goal previously stated is further subdivided in three smaller objectives

1. Automatic and robust speech segmentation at phoneme level: The FA system needs to be robust to errors such as repetitions, skipping of words or syllables, mispronunciation, insertion of superfluous phonemes.

2. Automatic Speech Recognition (ASR): It is hard to have a patient to read clearly and precisely a prompt. We formulate the hypothesis that the use of ASR in combination with FA could help in satisfying the robustness requirements stated above. Moreover, removing the need for audio recordings of known text/prompts allows analysis of spontaneous speech.

3. Leverage as much as possible unlabeled data: Easier to collect in large quantities, already available in relatively large quantities

1.2 Outline

This work is divided in three main parts

1. Introduction: this part includes a brief introduction to PD; an introduction to speech, its production, phonetics, phonology and prosody, speech disorders and voice pathologies, and the effects of PD on speech; a literature review for Automatic Voice and Speech Condition Analysis (AVSCA), ASR and FA; ending with a description of the material and overall methodology used in this work.

2. Experiments and solution implementation: describes the data preparations steps, the experiments and the results obtained.

3. Future work and conclusions: lists the contributions of this work, possible future developments and ends with the conclusions.

Chapter 2

Parkinson's disease

The brain represents less than 2% of a human body mass[172, 175] and yet it uses more than 20% of its energy production[56]. It contains about 86 billion neurons[11] which forms approximately 0.15 quadrillion connections[131, 191]. In NDs specific populations of these neurons, in the brains or in the peripheral nervous system, are characterized by progressive loss of function due to damage or cellular death[54].

NDs include AD, PD, Multiple sclerosis and others. AD is the most common NDs affecting 50 million people in 2020[25], followed by PD affecting 6.1 million individuals in 2016 [52].

NDs are a common and growing cause of mortality and morbidity worldwide, especially in the elderly[57]. Disease burden aggregates their impact on people, society and the economy in terms of mortality, morbidity, financial cost, and other factors; it is usually measured in terms of Disability-Adjusted Life Years (DALYs) or Quality-Adjusted Life Years (QALYs) and is reportedly increasing for NDs, due to increasing numbers of older people and potentially other environmental factors[52, 64].

More specifically the number of people with PD is expected to double between 2005 and 2030[139], while the global burden of PD has more than doubled, and it is predicted to increase substantially in the near future[52].

An early diagnosis might improve and maintain the quality of life of patients and increase their life expectancy, reducing the burden of these diseases. Currently, no reliable test yet exists to spot AD or PD before symptoms appear. Moreover, there are neither laboratory nor instrumental tests for monitoring the disease[50] that would be useful during treatment. It is worth pointing out that NDs share common characteristics and advances in one specific disease will probably benefit also other NDs.

In this chapter, we go on to discuss idiopathic PD, the most common type of disease categorized as parkinsonism, with the goal of providing an overview of this condition, then transition to how PD affects speech and how this could be used for its diagnosis.

2.1 Prevalence and incidence

PD incidence worldwide is estimated to be between 5 and more than 35 new cases per 100,000 individuals. The global prevalence for all ages is estimated at 0.3%, considering

only individuals 80 years old and older the prevalence increases tenfold to more than 3%[139].

In 2007 [51] predicted that by 2030, in the most populous nations, the number of PD patients could reach up to 9.3 million, more than double the maximum of 4.6 million individuals estimated to have PD in 2005. Values for 2016 from [52] seem coherent with this prediction reporting an estimate of 6.1 million individuals with PD globally. The same study reports also that the 74.3% increase in PD seen between 1990 and 2016 cannot be attributed solely to the increasing number of older people, which increased only by 21.7% in the same period. According to [97], other causes could be found in environmental factors (pesticides, herbicides, metals) and specific living conditions (rural living, farming).

2.2 Mechanisms

Parkinsonism results from a decreased dopaminergic transmission in the motor region of the striatum, a cluster of neurons with a critical role in the motor and reward systems.

This dopaminergic transmission to the striatum occurs through dopaminergic neurons from the substantia nigra.

The substantia nigra is part of the basal ganglia, located in the midbrain one of three parts of the brain stem, where the cerebrum connects with the spinal cord. It is divided into pars compacta and pars reticulata. The pars compacta supplies with dopamine the striatum.

The pathway from substantia nigra to striatum plays an important role in the control of motor functions.

Decreased dopaminergic transmission results from the two characteristic features that together are specific for PD:

- **Neuronal loss in specific areas of the substantia nigra**; in PD neuronal degeneration is limited to only certain types of neurons in particular brain regions. Initially limited to dopaminergic neurons in specific areas of the substantia nigra, later on in the disease becomes more widespread in the midbrain. Degeneration in this region has been suggested to start before the onset of motor symptoms.

- **Widespread accumulation of the α-synuclein protein**; abnormal deposition of α-synuclein in the cytoplasm[1] of certain neurons in several different brain regions. α-synuclein is a protein of yet unknown function, normally is mainly found in the nervous system, its presence should be dynamically regulated. The misfolding of α-synuclein is hypothesized to spread in a Prion-like fashion and is considered responsible for its agglomeration into Lewy's bodies. These agglomerations have been linked to dysfunction and death of certain populations of neurons[77]. Age and failing brain defenses are thought to have a role in the non-removal of excess α-synuclein[139].

[1] Gelatinous liquid that fills the inside of a cell[41]

How PD begins is still a source of discussion, the authors of [24] hypothesized in 2003 that PD begins when a foreign agent enters the body via the nose or gastrointestinal system and travels into the central nervous system through the vagus nerve. The same study proposed a division in stages of the disease progression based on neurological structures reached by neurodegeneration. More recently clinical and pathological evidence has been presented by [97] and others to support the hypothesis that PD starts in the gut, coherent with [24], reinforcing the possibility that environmental substances can trigger pathogenesis. On the other hand, [23] underlines the relative lack of post-mortem cases with isolated peripheral α-synuclein pathology for the moment, which weakens the case for a gastrointestinal onset, and that the pathology in most PD patients may originate in olfactory or gastrointestinal only or mainly because these are statistically the most likely sites of stochastic α-synuclein misfolding, excluding or limiting the importance of external stimuli.

2.3 Causes

So the causes of PD are not known yet. However, some risk factors have been revealed as influential in developing the disease, sometimes in relation to theories on possible mechanisms of the disease. The following categories of risk factors are usually identified, and generally affected individuals fall in a mix of them.

- **Genetic**: are estimated to contribute approximately 25% to the overall risk of developing PD[44]. Several genetic variants have been identified and each of them contributes a small amount individually to the risk of developing PD. The most impactful include the genes GBA, LRRK2, SNCA, PARK7 and PRKN. It is expected that even more genes will be found to impact the risk of PD.

- **Environmental**: over a dozen environmental factors have been associated with the risk of developing PD. Examples are: exposure to pesticides and use of specific pesticides, traumatic brain injury, exposure to high air pollution, gut microbiome conditions, metals, solvents[27]. It is believed that in most cases environmental factors are responsible to trigger PD pathogenesis and propagate it to the brain, the possibility that it starts in the olfactory and gastrointestinal systems agrees with this hypothesis[97]. Some examples of an inverse association with PD risk include caffeine consumption, smoking, vigorous exercise, and ibuprofen use. [27] however underlines that the newly discovered multiple systems' nature of PD coupled with decades-long disease initiation and prodromal development period make it very difficult to achieve reliable and valid exposure assessment in the most relevant periods for PD etiology.

- **Aging**: it remains the greatest risk factor, with age-related decline in midbrain dopaminergic neurons and decreased efficiency of defense mechanisms that would normally contrast protein accumulation.

2.4 Symptoms

Motor symptoms are those usually associated with PD, most common motor symptoms include

- Tremor
- Rigidity
- Bradykinesia or slowness of movements
- Postural instability
- Akinesia, that is a delay at the beginning of movements
- Hipokinesia, consisting of poor, incomplete or simplified movements (an example is smaller handwriting)

A variety of non-motor symptoms and signs has been identified too, including

- Depression and anxiety
- Constipation
- Olfactory disturbances or loss of smell
- Swallowing problems
- Communication problems
- Dementia, hallucinations, difficulties focusing and performing complex tasks
- Sleeping problems
- Sexual disruptions

Some of these non-motor symptoms like olfactory disturbances, sleep disorders, autonomic dysfunction (especially constipation) and depression can anticipate the onset of more visible motor symptoms by over a decade[97]. Vocal impairments have also been demonstrated by [156] to occur in 78% of untreated early PD subjects. These early appearing symptoms can play a key role in early diagnosis, which in turn can lead to an important quality of life improvement for individuals with PD.

2.5 Diagnosis and rating scales

The diagnosis of PD for now is still a clinical one and is not straightforward as one may think, due to different pathogenesis having similar and overlapping signs and symptoms. Clinical diagnosis requires follow-ups with continuous diagnostic re-evaluation to be accurate.

The diagnosis is now typically carried out following standard diagnostic criteria, the use of standard criteria improved the accuracy of the diagnosis and is now diffuse in medical organizations. The most known criteria comes from the United Kingdom Parkinson's Disease Society Brain Bank (UKPDSBB) and from U.S. National Institute of Neurological Disorders and Stroke (NINDS)[2]. The UKPDSBB was the first one to be published and uses three steps [91]:

1. Diagnosis of Parkinsonian syndrome: requiring the occurrence of bradykinesia and at least one of muscular rigidity, tremor or postural instability
2. Exclusion criteria for PD: in which the occurrence of signs common for other causes of parkinsonisms are used to possibly exclude PD as a cause
3. Supportive prospective positive criteria for PD: three or more supportive features among unilateral onset, rest tremor, progression of the disorder, persistent asymmetry, excellent response to levodopa, severe levodopa-induced chorea, levodopa response for 5 years or more, the clinical course of at least 10 years, are required for a definitive diagnosis

Ten years later in 1999 Gelb and other American co-authors published a new set of diagnostic criteria. Unlike UKPDSBB criteria, Gelb criteria were based on different levels of diagnostic confidence: possible, probable, and definite. The definite diagnosis only when PD is confirmed at autopsy [110].

An increased understanding of PD on different levels has led to the development, and publishing in 2015, of new diagnostic criteria from the Movement Disorder Society (MDS). This set of criteria proposed in [141], known as Movement Disorder Society (MDS) Clinical Diagnostic Criteria for Parkinson's Disease (MDS-PD), encompasses the two previously discussed sets of criteria and introduces new important aspects as the use of non-motor symptoms as possible diagnostic features and the adoption of the concept of prodromal PD, fundamental for research studies. For now, however, this criteria set still lacks pathological validation and is scarcely employed among clinicians[110].

Scales have also been defined to monitor the progression of PD in an individual, note that those should not be used to define or diagnose PD[141].

2.5.1 Unified Parkinson's Disease Rating Scale (UPDRS)

The Unified Parkinson's Disease Rating Scale (UPDRS) is the most commonly used scale for the assessment of parkinsonian motor impairment and disability[121]. The rating is accomplished through interviews with the patient and clinical observations. The original version[61] comprises four parts, scores are assigned for each part as follows:

1. mentation, behavior and mood: up to 16 points
2. activities of daily living: up to 52
3. motor: up to 56

[2] Also known as Gelb criteria, from the name of its main author

4. complications: up to 23

Each part is composed of items that can have yes/no answers or scored answers, with integers from 0 to 4 meaning none, mild, moderate, severe, marked respectively. The total score then goes from 0 to 147, where higher score mean higher affection from PD. The final score is a sum, so the same score can be obtained with the expression of different signs.

More recently MDS funded a revision (Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS))[72] of UPDRS. It introduces nine new items that were not captured in any form on the original scale: anxious mood, dopamine dysregulation syndrome, urinary problems, constipation, fatigue, doing hobbies, getting in and out of bed, toe-tapping, and freezing (objective rating). Yes/no answers have been changed to answers with values. The meaning of values has been changed from mild/moderate/severe/marked to slight/mild/moderate/severe, and the questions have been adapted accordingly, in order to better document changes in early stages of PD. Moreover, questions have been adapted so that for a large the patient or the caregiver can answer autonomously.

2.5.2 Hoehn & Yahr (HY)

The Hoehn & Yahr (HY) scale is simpler and easier to apply compared to the UPDRS[73]. It was introduced in 1967 in [86] as a system to grade the severity of PD into 5 stages corresponding to increasing levels of clinical disability. The broad definition of the scale stages does not permit consistent detection of effective interventions, this led to the introduction of intermediate stages such as 1.5 and 2.5[73]. For this reason, even if still widely used, it has largely been replaced as the primary measure of treatment efficacy by the UPDRS[73].

The modified HY scale as illustrated in [73] includes the following stages

- Stage 1: symptoms are very mild; motor signs are usually unilateral.
- Stage 1.5: unilateral and axial motor involvement.
- Stage 2: bilateral involvement without balance impairments.
- Stage 2.5: mild bilateral disease with recovery on pull test.
- Stage 3: mild to moderate bilateral disease; some postural instability; the patient is still physically independent.
- Stage 4: severe disability; still able to walk or stand unassisted.
- Stage 5: patient totally dependent. Wheelchair-bound or bedridden unless aided.

2.6 Treatment

PD currently remains an incurable disease, treatment has the purpose of slowing or halting the disease progression. Despite being diagnosed with the same disease, different

individuals with PD might have different symptoms with varied rates of progression leading to different treatment strategies.

Main treatments for PD are related to motor symptoms, typically beginning with pharmacological dopamine replacement; the most common drug is levodopa, usually mixed with carbidopa[28].

Levodopa provides the most significant symptomatic relief with least adverse side-effects. The addition of carbidopa helps in transporting levodopa to the central nervous system avoiding its conversion to dopamine in the peripheral tissues[28]. Main side effects for prolonged use of levodopa with carbidopa are dyskinesia[3] and fluctuating "off-on" periods of effectiveness[28]. To improve on the latter, methods of continuous delivery of levodopa/carbidopa have been developed, an example is Duodopa which administered in the intestine through a tube and a pump, and are under development[28].

To treat the same symptoms dopamine agonists are also used as an alternative to levodopa, early in the treatment due to the shorter duration of their effect, or jointly, to stabilize on-off periods[28]. They can be used with immediate or extended release. They can present problematic side effects related to impulse control[28].

Another alternative are Monoamine oxidase B (MAO-B) inhibitors, substances that inactivate the enzyme responsible for the inactivation of dopamine increasing the half-life of dopamine or levodopa in the brain[28]. They can be used as a monoterapy or in conjunction with levodopa[28].

For patients suffering from long-term complications from levodopa it is also possible to use Deep Brain Stimulation (DBS) targeting specific areas of the brain, this however requires a surgical operation to implant electrodes in the brain[28].

Rehabilitation programs targeted to help with effects of dopamine reduction such as softer voice and limits to body movements are typically recommended[28].

Moderate exercise has been found to improve the quality of life for people with PD and strenuous aerobic exercise likely has a neuroprotective effect[28].

[3]Involuntary muscle movements

Chapter 3

Speech

At the time of writing 7,151 languages are spoken[89], even though 42% of them are at risk of disappearing[88] in favor of other more prominent languages.

Speech is the primary realization of human languages, in fact every human language is or has been spoken at some time and thousands of them have only a spoken form[18].

Speech is characterized by the transmission of sounds and noises, from our phonatory apparatus that produces them, as sound waves through the air that are then received and perceived by a listener uditory system and brain.

Different types of speech production can be distinguished as

- *spontaneous*, like a conversation among few speakers in a familiar environment about a topic they choose by themselves
- *reactive*, when reading aloud a written word or naming a picture
- *imitative*, when an individual speaks the sounds they have heard another person pronounce

The production of speech is the process by which thoughts are translated into speech sounds, according to models that have been proposed in the last decades it incorporates the following stages, in some cases merging some of them into one[65]:

- conceptual stage, in which an abstract form of the proposition to be expressed is identified
- syntactic stage, where a frame, or sentence structure, is chosen
- lexical stage, in which words are searched based on meaning; once a word has been found information about its sense, possible collocations, phonology, and morphology become available
- phonological stage, where the abstract information collected until now is converted into a speech-like form
- phonetic stage, when instructions for the muscles that control the articulators are prepared to be sent, actual motor planning and actuation/articulation

What follows is a short introduction to phonetics, with a focus on speech production, a brief overview of phonology and morphology. Finally, the chapter ends with a section on voice disorders and a section on the effects of PD on speech.

3.1 Phonetics: articulatory and acoustic

Phonetics is the branch of linguistics that studies speech sounds, it is further subdivided in three sub-disciplines:

- *articulatory phonetics*, which studies how human produces sounds
- *acoustic phonetics*, which deals with acoustic characteristics of sounds, some examples could be amplitude, duration, waveform and fundamental frequency
- *auditory phonetics*, which focuses on hearing and perception

In the context of this work it is important to introduce some concepts from the first two sub-disciplines, they are therefore detailed in the next subsections.

3.1.1 Biological aspects of speech and phonatory mechanism

Not one of the organs used for speech production neither any of their anatomical aspects can be singled out as specialized for the purposes of producing speech yet[100]. Speech production can thus be considered a function overlaid on more primary biological system, its actions exploit neuromuscular capabilities with different primary biological functions such as breathing, sucking, biting, chewing, swallowing, licking, spitting, sniffing, clearing the throat, coughing, yawning, laughing, crying, shout to threaten[100].

The fact that a listener can hear speech depends on the presence of a moving stream of air. Language sounds are normally produced with an egressive air flow (sounds produced with an ingressive flow or with no flow also exist and are called ”avulsive”). Air is expelled in a carefully controlled fashion at a slower rate than normal breathing. The expulsion is controlled by the action of the muscles moving the diaphragm and by the various pairs of muscles that act on the rib cage. The air stream is formed in the lungs, there from the bronchi moves up into the trachea reaching the larynx, where it meets the vocal folds, and then procedes to the supraglottal cavities, that are the pharynx (or throat), the oral cavity (mouth) and nasal cavity (nose).

The larynx is the structure composed by vocal folds and their cartilaginous housing. The larynx could be described as a ”box” made of cartilages in which the frontal wall would be the thyroid cartilage (responsible for the ”Adam’s Apple”) in the front of the neck. Eight other cartilages are found in the larynx (for a total of nine): the cricoid cartilage (almost ring-shaped in the lower part of the larynx where it connects with the trachea), the epiglottis (spoon-shaped, located in the upper part of the larynx, closes access to the larynx when swallowing), a pair of arytenoid cartilages (triangular shaped, located in the posterior part of the larynx, they influence the position and tension of the vocal folds), a pair of corniculate cartilages (small and horn shaped, located on the top tips of the arytenoid cartilages), and a pair of cuneiform cartilages (elongated shape, they

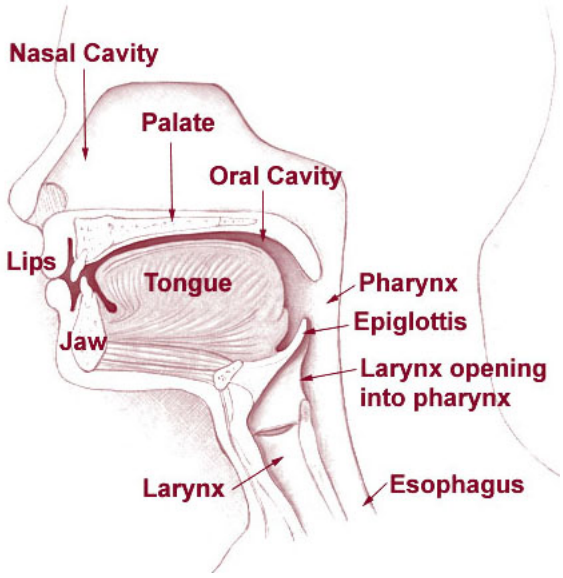

Figure 3.1. Head and neck

sit on top of the arytenoid cartilages, they support the vocal folds). The muscles in the larynx are classified in extrinsic and intrinsic. Extrinsic muscles connect to the larynx and other parts around it, they support and position the entire larynx. Intrinsic muscles are the ones inside the larynx and are responsible for the movement of the vocal folds, for their abduction (opening), adduction (closing) and tension. There are six intrinsic muscles:

- the cricothyroid muscle responsible for elongation and tension of the vocals folds
- the posterior cricoarytenoid muscles open the vocal folds for respiration
- the lateral cricoarytenoid muscles oppose the posterior cricoarytenoid muscles, they close the vocal folds to protect the airway
- the transverse arytenoid muscle closes the vocal folds/glottis especially in the back part
- the oblique arytenoid muscles and the thyroarytenoid muscles work together to open or close access to the larynx when swallowing or coughing

All these muscles, except the cricothyroid, are innervated by the recurrent[1] laryngeal nerves, a branch of the vagus nerve[157]. The cricothyroid muscle in innervated by the superior laryngeal nerve, another branch of the vagus nerve[157].

[1]It is called recurrent because it extends in a direction opposite to that of the nerve it branches from, specifically in this case it loops around the aorta before going back to the larynx; this happens also in other vertebrates, the most remarkable case is in the giraffe, where it extends for about 4.5 meters[31]

The vocal folds are two specular infoldings of mucous membrane, they stretch horizontally back to front attaching anteriorly to the thyroid cartilage and posteriorly to the anterolateral surface of the arytenoid cartilages[190]. They are composed of three different tissues stacked one on top of the other: the upper one is a non-keratinized squamous epithelium, under it is located a gel-like layer called lamina propria allowing the vocal folds to vibrate, and under that are found the vocalis and thyroarytenoid muscles. The vocal folds are separated and relaxed during normal breathing, but they can contract and stretch to reduce or block air flow leading to their vibration during phonation. The space between the vocal folds can thus be free or partially free, and in these cases is called glottis, or completely blocked. Their size affects the pitch of voice, they are about 11–15 mm long in adult women and 17–21 mm in men[190].

The vestibular folds, also known as false vocal folds, are located on top of the true vocal folds. In contrast to the true vocal folds they have a minimal role in phonation, limited to deep sonorous tones, screams, and growls, they protect the more delicate vocal fold allowing or closing access to the larynx.

After flowing between the vocal folds the egressive pulmonic air stream enters the pharynx (throat), from there it goes on in the oral cavity and, if the velum allows it, to the nasal cavity. In the oral cavity active (mobile) and passive (fixed), organs contribute to speech production.

The pharynx is above the trachea and the oesophagus (the tube going down to the stomach) and behind oral and nasal cavity. Here food and air entering the human body are directed by the epiglottis, a flap of cartilage, respectively toward the stomach through the oesophagus and the lungs through the larynx and the trachea. Due to its position it plays a role in voicing and articulation.

The oral cavity is found in front of the pharynx. It is bounded in the frontal part by the lips and teeth on the alveolar bone. In the top by the palate (hard in the front, soft in the back). The floor is formed by the mylohyoid muscle (running from the neck to the mandible) and the tongue. A mucous membrane encloses the sides. The uvula projects downwards from the soft palate. Many articulators are located in the oral cavity.

The tongue constituting the floor of the mouth is a muscular organ covered in moist mucosa, it starts with the two unattached portions of the tip (technically caled apex) and the blade, its anchored part is only partially visible and extends downwards where its root forms the front wall of the pharynx[10]. It is involved in taste, swallowing, digestion and cleaning of teeth. Its importance in speech is due to its extreme flexibility due to the fact that its parts can move relatively independently, to the point that they are considered separately in phonetics.

The nasal cavity, above the oral cavity, begins at the velum and ends at the nostrils. There are actually two nasal cavities, one extending from each nostril, each divided in two segments, one in the bottom with respiratory purposes and one in the top for olfactory purposes, where the olfactory nerve is located.

3.1.2 Phonation

Phonation, or voicing, is the process characterized by rapid openings and closings of the vocal folds which leads to their quasi-periodic vibration and production of sound.

Although there have been different explanations for what makes the vocal folds

vibrate, it is now generally agreed that the basic theory providing the answer is the myoelastic-aerodynamic theory of phonation[192]. Formulated by van den Berg[170] and later revised in [165, 192], its description of the mechanism of vocal folds self-oscillations during phonation can be summarized as follows[192]:

- the exhalatory airflow starts and its blocked below the closed vocal folds, air pressure higher than above the folds builds up
- this pressure pushes the lower margins of the vocal folds apart, starting to open the glottis; the upper part of the folds opens with a delay
- the elasticity of the vocal folds tissues leads them to start closing, the tenser the vocal folds the sooner they start closing (leading to a higher fundamental frequency as will be soon discussed)
- the lower margins starts closing before the upper ones
- this causes the folds to form a shape that leads to the drop of intraglottal pressure, this pulls the vocal folds together
- due to their elasticity the movement reverses and the glottis starts opening again from below, and the cycle repeats

This cycle repeats on average 100-120 times per second for adult male voices and 200-240 times per second for adult female voices, in children it can be as many as 300 times per second[10]. These rates can be expressed as frequencies in Hertz, more specifically they all describe the Fundamental frequency (F0). The different rates or frequencies of vibration are heard as pitch, a variation introduced going up or down in pitch adds intonation. Some languages use intonation to make distinctions between different words and are called tone languages, other languages use an intonational phrase using tone to make a phrase sound more like a question or a statement, or make it seem incomplete.

The oscillation of the vocal folds is mostly lateral and partially toward the top, it serves to modulate the pressure and flow of air through the larynx. This modulated airflow is the main component of most voices phones. In phonetics the word voice refers to sounds produced by vocal fold vibration, or *voiced* sounds, in contrast to *unvoiced* or voiceless sounds which are produced without or with very little vocal fold vibration. There is no vocal folds vibration when the folds are placed laterally (abducted) and are not close enough to each other, when they are too much or not enough tensed, or where the pressure drop is not big enough to encourage their closing after their opening. In the next sections the voiced/unvoiced parameter in conjunction with two other parameters, that will be introduced later, are used to construct a label for each phone.

Different vibratory patterns are possible and give rise to different vocal registers, range of tones in human voice corresponding to a specific pattern. Some examples are:

- modal
- breathy voice
- creaky voice

From an acoustic phonetics point of view, speech is a complex periodic wave that can be decomposed into the sum of several sinusoidal signals, each characterized by a specific frequency and intensity (according to Fourier's Theory). All frequencies components included in the signal spectrum are integer multiples of a frequency, namely F0, which is responsible for the pitch perception of speech[95, 99]. The multiples of F0 are in turn called *harmonics* or harmonic overtones. This complex periodic wave is also called Larynx waveform (Lx), its period has a triangular looking shape with a gentler angle in the opening phase, indicative of the exertion required to push the folds apart (meaning the movement is slower, occupying more time), and a sharper angle in the closing phase, indicating a much quicker drawing back together of the folds when the sudden drop in pressure occurs[10].

An interesting way of viewing voice production is that proposed in source-filter theory[62], which sees the speech production mechanism as composed of two stages[167]:

1. the above described vibration of the vocal folds due to air coming from the lungs, that produces Lx composed by F0 and its harmonics, that is the *"source"* sound

2. the air and the sound wave goes up in the vocal tract where frequency components are amplified or diminished based on the resonances of the vocal tract, that acts as a "filter", based on the position of its organs (places of articulation), that will be discussed more deeply in the next section

This selective modification of the voice source spectrum produces perceptible contrasts, which are used to convey different linguistic sounds and meaning[190].

3.1.3 Articulation

The process of articulation consists in the movement of speech organs, normally two, towards each other to contact and create an obstruction. The goal is to "shape" the air flowing out of the larynx to obtain sound quality useful for speaking.

The variables relevant for this process are two

1. the "place of articulation", that is the point where organs meet to create the obstruction

2. the "manner of articulation", that is the way the obstruction is formed and released

Going more into details, a place of articulation is usually characterised by a fixed point in the vocal tract, called also a *passive* or *stationary* place of articulation, that is approached or contacted by its relevant active organ, also called *active* or *mobile* place of articulation. For the greater part passive organs are located along the upper surface of the vocal tract while corresponding active articulators are along the lower part, see figure 3.2.

The *place of articulation* is used in conjunction with *manner of articulation* and presence or absence of voicing for classifications of phones. What follows is a list of place of articulations with their corresponding adjectives, used in classification:

- active places of articulation, usually in the lower part

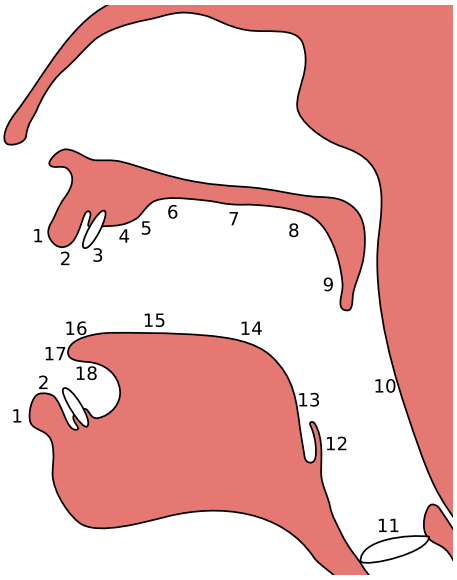

Figure 3.2. Places of articulation: 1 Exo-labial, 2 Endo-labial, 3 Dental, 4 Alveolar, 5 Post-alveolar, 6 Pre-palatal, 7 Palatal, 8 Velar, 9 Uvular, 10 Pharyngeal, 11 Glottal, 12 Epiglottal, 13 Radical, 14 Postero-dorsal, 15 Antero-dorsal, 16 Laminal, 17 Apical, 18 Sub-apical

  - lower lip (labial)
  - various parts of the tongue: due to the previously described flexibility of the tongue it is further subdivided in
    * front (coronal), further divided into
      · tip of the tongue (apical)
      · blade (laminal), just behind the tip
      · under the tip (subapical)
    * body (dorsal)
    * base/root (pharyngeal)

    another parameter related to the tongue is the shape it assumes during articulation
    * central
    * lateral
  - aryepiglottic fold (aryepiglottal)
  - glottis (glottal)
- passive places of articulation, usually in the upper part
  - upper lip (labial)
  - upper teeth (dental)
  - alveolar ridge (alveolar)

  - back of the alveolar ridge (post-alveolar)
  - hard palate (palatal)
  - soft palate (velar)
  - uvula (uvular)
  - pharynx (pharyngeal)
  - epiglottis (epiglottal), at the entrance of the windpipe

In the same way, a way to classify manner of articulation is presented:

- obstruents
  - plosives, also called stops, are characterized by an occlusion in the oral vocal tract and no nasal air flow
  - fricatives, characterized by continuous turbulent and noisy airflow produced by air forced through a narrow channel obtained placing two articulators close together
  - affricates, begin like plosives and end like fricatives
  - stridents, containing some fricatives and some affricates
    - sibilants
- nasals, characterized by an occlusion of the oral tract with air passing through the nose
- vowels, characterized by no obstruction in the airflow and the presence of phonation
- approximants, characterised by very little obstruction
  - semivowels, like a vowel but with the tongue closer to the roof of the mouth
- vibrants
  - trills, when the airstream causes an articulator, held in place, to vibrate
  - taps/flaps, a momentary closure of the oral cavity
- liquids
  - laterals, the airstream proceeds along one or both of the sides of the tongue, but it is blocked by the tongue from going through the middle of the mouth
  - rhotics

For classification and labeling phones are divided in two macro categories[18]

- consonants, can be voiced or voiceless, are characterized by the presence of obstacles partially or completely obstructing the airflow along the tract between the glottis and the end of its path; they are labeled based on 3 parameters: voiced/voiceless, place of articulation, manner of articulation

- vowels, characterized by the presence of phonation and the absence of obstacles obstructing the airflow; they are classified based on the different configuration assumed by the vocal cavity for their production, in particular the position of the tongue

The parameters for the classification of consonants have already been discussed, are now introduced the parameters needed for vowels classification.

Vowels articulation is described mainly in function of tongue position and shape of the lips. Main articulatory features of vowels, useful for classification, are

- *height*, theoretically referring to the vertical position of the tongue (or the jaw, depending on the model), goes from high (or close) to low (or open) in seven steps/degrees; it influences the first formant First formant (F1) (first/lowest harmonic of F0): the higher the frequency of F1, the lower (more open) the vowel
- *backness*, theoretically referring to the horizontal position of the tongue relative to the back of the mouth, it is specified in five steps from "front" to "back"; it influences the second formant Second formant (F2): a "front" position corresponds with a high frequency F2, the opposite is true for a "back" position
- *roundedness*, from the rounding of the lips, we may distinguish rounded and not rounded vowels; rounded vowels are characterized by a lower F2 and slightly lower F1 compared to their not rounded counterpart

3.2 Phonology

Phonology is the branch of linguistics that studies how languages organize their sounds.

3.2.1 Transcriptions, phones, phonemes and allophones

Often the written form of a sentence or a word in a given language has not a corresponding pronunciation easily derivable trough rules. In Italian for example is enough to learn a few rules to read sentences and sound relatively authentic even without knowing the language itself. The same cannot be said about English for example.

The rules referenced in the previous paragraph are needed to translate characters or better *graphemes*, the smallest functional unit of a writing system, to the equivalent units in sounds of a given language. For a particular language, the unit of sound that can distinguish one word from another is called *phoneme*. A phoneme can be considered an abstraction, it represents a set of sounds that in a particular language are not distinguishable or, in another way, are perceived equivalently. Note that sounds that are perceived to be the same phoneme in a language can be distinguished and perceived as different phonemes in another.

The unit of sound in general, without considering any specific language, representing every possible sound that the phonatory system can make, is called *phone*. When two phones can be used to realize the same phoneme, in a given language, they are said to be *allophones*.

In the same way as graphemes have a concrete written representation, there exists different alphabetic systems to associate a symbol to phone or phoneme. Examples could be

- International Phonetic Alphabet (IPA), most widely used and well known, its 2020 revision is illustrated in figure 3.3
- Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA), designed to map the 1993 IPA into 7-bit ASCII

Other alphabetic systems also exist, some of them were developed primarily to be ASCII-based encoding of IPA before it became more usable thanks to wide availability of Unicode, others are instead language specific.

Symbols from an alphabetic system for phonetic notation can be used to write transcriptions, which in turn can be

- narrow, distinguishing a specific phone in a set of allophones or encoding other useful information to reproduce exactly a specific sound
- broad, when only the most noticeable phonetic features are transcribed, an example of broad transcription is the phonemic transcription (which disregards allophonic differences)

3.2.2 The phonemes of the Italian language

Different languages have a different phonemic inventory, set of distinctive sounds, of different sizes. There exist spoken languages with as few as 11 phonemes in their phonemic inventory and others with up to 140[18], the number of phonemes for a single language changes depending on the convention used to count them. Standard Italian counts 30 phonemes, but depending on whether one does not consider separately semivowels or considers separately consonant lengthening the count becomes 28 or 45[18].

Of the 30 phonemes of the standard Italian 7 are vowels (see table 3.1) and 23 are consonants (table 3.2).

	Front	Central	Back
Close	i		u
Close-mid	e		o
Open-mid	ɛ		ɔ
Open		a	

Table 3.1. Vowels of the Italian language[3, 18, 19, 80]

3.2.3 Syllables

Syllables are the “building blocks” used to compose the sound of words, a syllable is a minimum combination of phonemes that can be pronounced[18]. They are the smallest unit for suprasegmental and prosodic facts[143].

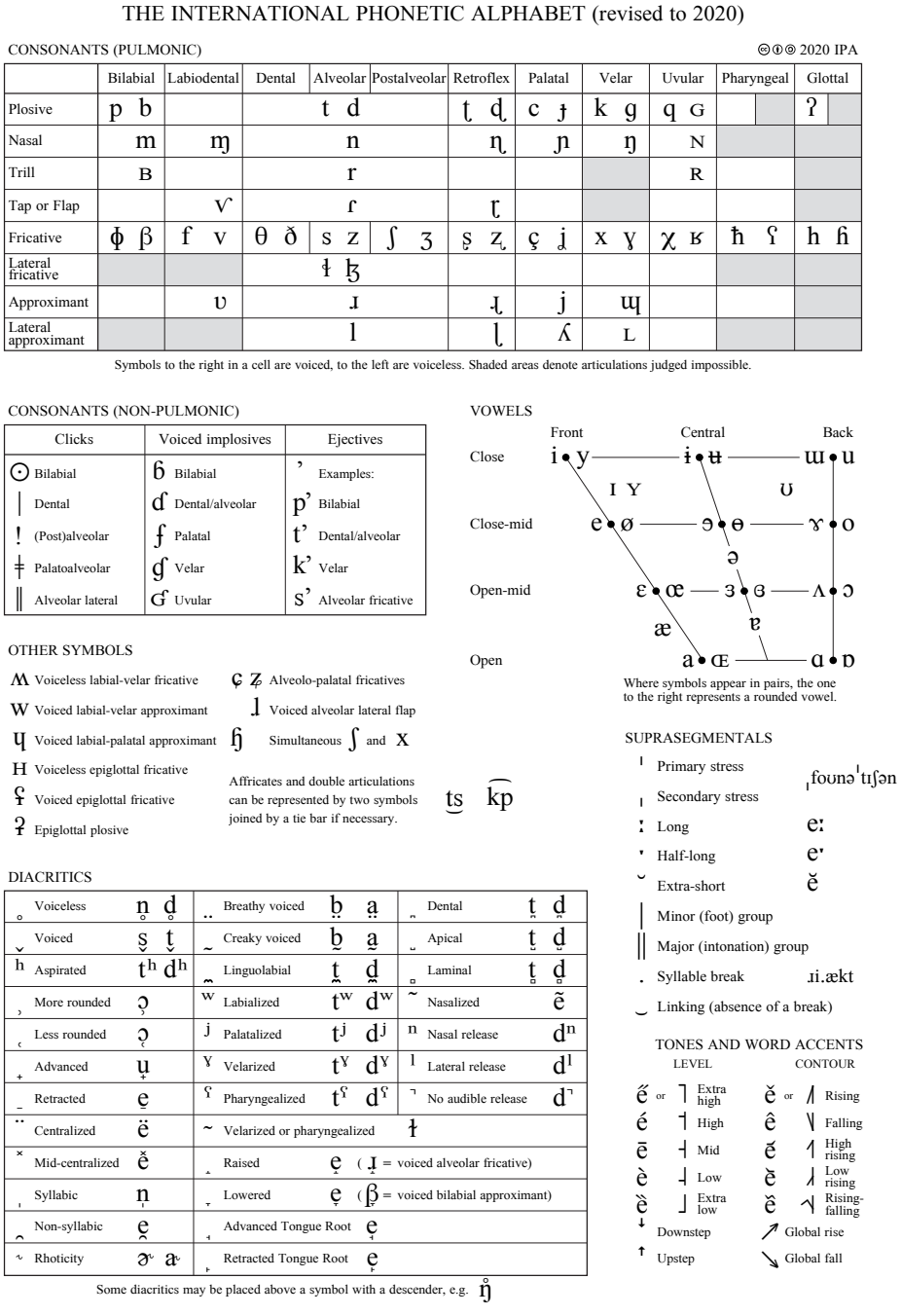

Figure 3.3. Full IPA chart (2020 revision) from [92]

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Palatal	Velar
Plosive	p b		t d				k g
Affricate			ts dz		tʃ dʒ		
Nasal	m			n		ɲ	
Trill				r			
Fricative		f v	s z		ʃ (ʒ)		
Approximant						j	
Lateral approx.				l		ʎ	

Voiced labial-velar approximant | w

Table 3.2. Consonants of the Italian language[3, 18, 19, 80], symbols to the right in a cell are voiced, to the left are voiceless. The phone in parentheses does not belong to the Italian phonemic inventory by itself, but only in its affricate forms.

In acoustic terms there are two main theories defining the syllable: one in in terms of sonority the other in terms of prominence[143]. According to the first theory the syllable is a portion of word between two minimums in sonority[113], the (relative) peek in sonority in-between is defined as the nucleus of the syllable[143]. Every spoken word or sentence is thus constituted by an alternating sequence of phones that have higher sonority, corresponding to a more open vocal tract, and phones with lower sonority, corresponding to a more obstructed vocal tract[113].

In most languages, Italian included, a syllable is built around a vowel[18, 143]. Consonants and semivowels are added around the vowel that in a syllable is called "nucleus", while the consonants and semivowels added are called "onset" when before the nucleus and "coda" after the nucleus[18, 113, 143]. The term "rhyme" is used to refer to the set of nucleus and coda[143]. The onset and the coda can be composed by more than one element[113]. In Italian usually syllables have an onset composed of zero to three consonants (a-more V, ca-sa CV, pla-nare CCV, strap-po CCCVC, where C represents a consonant and V a vowel) and a coda composed by zero or one consonant[143]. Syllables that end with a coda, that is with a consonant or semivowel, are called "closed" while those that don't are called "open"[18, 113].

In other languages, it is possible for lateral, vibrant or nasal consonants to be the nucleus of a syllable, due to the relatively higher sonority compared to other consonants[113].

When one or two semivowels are present in a syllable together with a vowel respectively a diphthong or a triphthong is obtained[18]. A diphthong, also called gliding vowel, is a vowel involving the movement of the tongue between two different target positions, one for the first vowel sound and one for the second in the same syllable[10]. A triphthong is always a monosyllabic vowel combination involving a quick but smooth movement of the articulator from one target position, to a second, and then to a third.

3.3 Prosody and non segmental aspects

From the Greek "pros" (towards) and "ōidē" (song), the word prosody originally meant "song sung to music"[147]. In modern phonetics this word is instead related to the properties of speech that cannot be derived from the segmental sequence of phonemes underlying human utterances[126].

A segment is a distinguishable sound in a specific language[10], a phone (vowel or consonant) composing a word or a phrase[29]. Non-segmental aspects are concerned with what happens when phones contact and phenomena involving units larger than phones, they are subdivided accordingly in intersegmental and suprasegmental.

The main intersegmental aspects are coarticulation and phonotactics[113].

Suprasegmental facts are related to syllables, stress, tempo, rhythm and intonation[29, 113].

3.3.1 Connected speech and intrasegmental aspects

The term connected speech is used to refer to sounds forming words or sentences in spoken language when analysed as a continuous sequence[37]. Analysis of connected speech, as opposed to linguistic units seen in isolation, shows sound changes affecting phrases, words, lexemes, morphemes, syllables, phonemes or phones[37].

In the articulation of a phone three phases are typically distinguishable[143]:

- *onset phase*, when the articulators move to reach the appropriate position for the production of a given sound
- *medial phase*, in which the vocal tract maintains the optimal configuration it has reached for sound production
- *offset phase*, when the articulators move away from the current configuration and possibly prepare for the configuration needed for the next phone

When phones are inserted in connected speech they are usually influenced by preceding and following phones[113]. The movements that the phonatory organs have to make to reach the required positions for a given phone change depending on the position in which they were for the previous phone and on the position in which they need to be for the next phone[113].

This is seen as an overlap between onset and offset phases for consecutive phones, and it gives origin to mutual interferences named *coarticulation*.

In some cases coarticulation can lead to more prominent effects, causing the involved phone to transform in a different phone, these cases are described as assimilation and are common in many languages[113].

Coarticulation and assimilation are further subdivided in two types

- *anticipatory*, more frequent, when phonatory organs are taking the required position for the articulation of a particular phone and an articulator not involved in it begins to move in the direction of an articulation needed for a later sound in the utterance[37, 143]

- *perseverative*, when a sound retains a characteristic deriving from an earlier articulation[37], because movements or positions from a previous phone are kept in the articulation of the current phone[143]

An intersegmental aspect affecting coarticulation is *tempo*[113], that is the speed of speaking, also known as *rate*[37]. It can be measured in phones per second, it is usually described with the terms "allegro" and "lento", borrowed from music, respectively meaning faster and slower than usual[37, 113].

Tempo can affect speech making it[113]

- *ipoarticulated*, usually associated with a faster speed of speaking, is a characterized by prominent coarticulation leading to lower articulatory precision, simplification of sequences of vowels or consonants, weakening or omissions of phones or syllables. Some speakers are still able to properly articulate speech also at higher speeds, while others are characterized by ipoarticulation also at lower speaking speeds

- *iperarticulated*, usually associated with slower speech, typical of reading aloud or dictation. In iperarticulated speech, phone weakening and coarticulation are present only when "mandatory" for a specific phones sequence

Another intersegmental aspect is related to phone sequences and is called *phonotactics*. More precisely phonotactics concern the sequential arrangements of phonological units which occur in a language, that is: what counts as a phonologically well-formed word[37] based on a set of principles that limit the possible successions of phones[113]. As already seen in 3.2.3, only some alternances of consonants and vowels are allowed in Italian, similar "rules" are present also in other languages. In addition to that, only some sequences of specific consonants can be used, to form syllables, in a given language[18].

3.3.2 Suprasegmental aspects

Effects which extends over more than one sound segment, that is more than one phone[37]. As previously stated, suprasegmental aspects include *length*, *stress*, *pitch* and *rhythm*, that are now described a little more in detail.

Length refers in phonetics to the duration of a sound or utterance (in this case is also simply called duration) and in phonology to the relative durations of sounds and syllables (in this case is also called quantity)[37]. In phonetics length is defined as the time between the onset of a phone and its offset[18], in theory vowel articulation can be hold and thus last indefinitely[18] but, in Italian, short ones usually last 40-80 milliseconds while long ones are around 80-150, their length depends on syllabic structure[143]. Fricative consonants can in theory be held indefinitely too, while other consonants can be held only momentarily[18]. In phonology long and short lengths are recognized both for vowels and consonants, that is languages often have one degree of phonological length, and may have more than one[37]. For example long vowels occur as distinguishable phonemes in Arabic and Finnish but not in Italian or French in which they are allophones, while long (or double) consonants occur and are distinguishable phonemes in Italian and Luganda but not in English or French[18, 37, 143]. The long and short adjectives should be considered only in relative terms[18].

The term *stress*, or accent, in phonetics refers to the force or intensity used in producing a syllable[18, 37] in comparison with other syllables in a word or phrase[18]. Stressed and unstressed syllables are usually distinguished, the former being more prominent than the latter and marked in transcription with a raised vertical line[37]. The prominence of the stressed syllable is usually due to an increase in loudness, but increases in length and often pitch may contribute to the overall impression of prominence[37]. The position of the stress in a language can be fixed or free, in the former case its position its predictable[10, 18]. Some pairs of words or word sequences in some languages (Italian included) may be differentiated only by stress variation[18, 37].

Pitch is the acoustic correlate of the frequency with which the vocal folds vibrate[143]. The *intonation* is instead the melodic trend with which a phrase or an entire tonal group or rhythmic group is pronounced, that is how pitch varies while speaking[18]. Intonation in most languages allows distinguishing, for example, among affirmations, exclamations or questions[18].

Rhythm in this case refers to the division of time into equal portions by languages, called also "isochrony"[10]. Languages can be syllable-timed, mora-timed or stress-timed[10].

3.4 Speech and language disorders, voice pathologies

In 2000, [154] reported that communication disorders were estimated to have a prevalence of 5% to 10%. Using data collected in the same time period [153] reported that the lifetime prevalence of a voice disorder was 29.9%. Voice disorders belong, with fluency and articulation impairments, to the category of speech disorders, which in turn are a subcategory of communication disorders[46].

A communication disorder is defined by [46] as "an impairment in the ability to receive, send, process, and comprehend concepts or verbal, nonverbal and graphic symbol systems". Communication disorders are caused by one or a combination of speech disorders, language disorders, hearing disorders and auditory processing disorders[46]. A communication disorder may result in a primary disability, or it may be secondary to other disabilities; it may range in severity from mild to profound, and it may be developmental or acquired[46].

This chapter provides a brief introduction to speech and language disorders in general and then briefly elaborates on their incidence in NDs and more specifically in PD.

3.4.1 Speech disorders and voice pathologies

A speech disorder is defined as impairment in any of its inter-related components[43], are considered subcategories of speech disorders[46]:

- *articulation disorders* consist in atypical speech sounds production characterized by substitutions, omissions, additions or distortions that may interfere with intelligibility; an example could be difficulty producing [r] sounds in the respective language's standard pronunciation; articulation disorders are not to be confused with motor speech disorders, such as dysarthria (in which there is actual impairment of

the speech musculature) or developmental verbal dyspraxia (in which motor planning is severely impaired)

- *fluency disorders* are a change or interruption in the flow of speaking characterized by atypical rate, rhythm, and repetitions in sounds, syllables, words, and phrases

- *voice disorders* are characterized by abnormal production and/or absences of vocal quality, pitch, loudness, resonance, and/or duration, which is inappropriate for an individual's age and/or sex

As previously stated, a voice disorder occurs when voice quality, pitch, and/or loudness differ or are inappropriate for an individual's age, gender, cultural background, or geographic location[174]. It can be hard to distinguish what is appropriate or normal and what is not because of different perceptions of what is normal and factors that affect what should be considered normal. However, according to [79], it is possible to identify some common characteristic that can be regarded as synonyms of non-pathological voice condition. Specifically they refer to the perceptual definition presented in [9], that can be summarized in the following points

- a pleasant quality, with an absence of noise, inappropriate breaks, perturbations or atonality;

- pitch in accordance to the age and sex of the speaker;

- loudness that is appropriate to the communication event;

- pitch and loudness variations that are available to express emphasis, meaning or subtleties indicating individual feelings and semantic differences;

- sustainability to meet social and occupational needs.

Abnormal voices do not possess any, a combination, or all of the above properties. Typically, three types of aberrant voices are usually identified[9, 79]:

- muteness, characterized by the absence of vibration of the vocal folds coupled with the inability to produce audible sounds

- aphonia, characterized by the absence of vibration of the vocal folds and the ability to produce sounds, resulting in a voice described as extremely breathy

- dysphonia, described as the absence of vocal quality, pitch, loudness, and/or variability appropriate for an individual's age and/or sex

[79] proposes to consider superclasses "vocal aspects", comprising pitch and loudness, and "vocal quality", comprising resonant characteristics and vibrational patterns of the focal volds. Vocal aspects are seldom studied, much more attention has been dedicated to vocal quality[79].

3.4.2 Language disorders

A language disorder is impaired comprehension and/or use of spoken, written and/or other symbol systems[46]. The disorder may involve in any combination[46].

- Form of language
  - phonology, the system of sound for a language and the rules that govern sounds combinations in it
  - morphology, structure of words and construction of word forms
  - syntax, order and combination of words to form sentences, relationships among the elements within a sentence
- Content of language
  - semantics, meanings of words and sentences
- Function of language
  - pragmatics, combines the above language components in functional and socially appropriate communication

3.5 Effects of Parkinson's disease on speech and discourse

Voice disorder in PD is prevalent affecting 70 to 90% of those with PD, abnormalities in acoustic analysis are found as frequently even in early PD[108].

Over 200 years ago, in "An Essay on the Shaking Palsy"[134], James Parkinson described motor and non-motor symptoms of PD. Most of the essay was dedicated to motor problems, especially tremor, however several non-motor problems were mentioned too, like sleep problems and constipation[43]. Notably, speech and swallowing impairments in the advanced disease are described as "His words are now scarcely intelligible and he is not only no longer able to feed himself, but when the food is conveyed to the mouth, so much are the actions of the muscles of the tongue, pharynx and impeded by impaired action and perpetual agitation"[43, 134].

3.5.1 Why PD affects speech?

Emerging evidence suggests that voice dysfunction is the earliest sign of motor impairment in PD[108], a way to justify this is that fine motor control involved in vocalization probably results in dysfunction before the limbs, reductions in the variability of F0 of speech can be noticed up to 10 years before diagnosis[85, 140].

3.5.2 Characteristics of PD speech

Speech abnormalities occur across the interconnected domains of[127]

- phonation, with occurrence of reduced voice volume (hypophonia) and altered voice quality (dysphonia)

- articulation, which is impaired by a reduction in the range of articulatory movements (hypokinetic articulation)
- prosody, with the manifestation of dysprosody by flattened pitch inflection (monopitch) and loss of stress (monoloudness)

PD speech features include also festination and hesitancy.

3.5.3 Treatment

Treatment modalities include[43]

- Medication optimization or pharmacological intervention,
- speech therapy,
- surgical treatment, such as
  - deep brain stimulation (DBS)
  - vocal fold augmentation
  - others...

Chapter 4

Literature review

This chapter presents a literature review on the topic of automatic voice and speech analysis for the diagnosis of specific medical conditions, with a focus on PD, covered by section 4.1. Later, sections 4.2 and 4.3 provide a review of key publications and recent advances in the fields of ASR and FA respectively and conclude by analyzing their use in the context of 4.1.

4.1 Automatic voice and speech condition analysis

The complex nature of speech production leads to a product that embeds much more data than the linguistic content encoded in sounds it was meant to convey[78]. This embedded data is usually described in four dimensions[78]

- *paralinguistic*, information about the affective, attitudinal or emotional state of the speaker
- *extralinguistic*, identity and state of the speaker such as age, sex, condition and similar
- *linguistic*, the message, variations in language, dialect, sociolect, idiolect and speech style
- *transmittal*, physical location of the speaker

Many publications exist in the literature for automatic systems leveraging speech to isolate a component from these dimensions, examples are systems able to determine speaker identity, age, sex, emotions, level of interest, accent, and dialect[78].

In the clinical setting, the use of voice recordings for the analysis of speaker condition is gaining popularity[78]. For example, acoustic analysis of the subject's speech by an expert clinician is part of the diagnosis and monitoring of many common neurological conditions, like PD[15, 125].

Traditionally, the diagnosis of speech or voice disorders can involve neurological, radiological, psychological, instrumental and perceptual (acoustic) assessment methods[15, 78]. The latter one extracts not quantifiable multidimensional information and

describes them qualitatively[78]. Moreover, during a typical outpatient visit, the clinician has to interview the patient to identify any symptoms, examine previous medical records, and perform a careful investigation of the overall health condition[78].

This process is[15, 78]

- time-consuming and expensive, leading to less frequent evaluations and more inconsistent results due to a high dependence from the patients' conditions at the time of recording;
- subjective, given that the evaluation is strongly dependent on the clinicians' expertise

To mitigate these problems, AVSCA[1] systems have been developed to analyze, classify, and quantify voice alterations consequent to speech disorders. Given the close relationship that exists between acoustic features extracted from speech or voice and related pathology, automatic systems can and have been designed to provide objective measurements of a patient vocal condition, reducing the evaluation time and the cost of diagnosis and treatment[78]. They also provide the added advantage of avoiding invasive procedures, employing speech signals which are easily recorded by inexpensive means[78].

AVSCA systems have been extensively applied to laryngeal pathologies, several examples exist also for other disorders such as PD, obstructive sleep apnea, AD, dysphagia and lupus[78]. Speech analysis has been singled out as a cost-effective and reliable method for detecting the presence of mild cognitive impairment[111] and NDs such as AD[111] and PD[118]. In the past ten to twenty years, many articles and studies have been published identifying characteristic speech features useful for the accurate distinction of healthy elder people and those affected by PD[119, 125] or AD[111].

In this context, 4.1.1 introduces the building blocks of a AVSCA system, 4.1.2 describes difficulties in the actual use of AVSCA systems, while 4.1.3 presents applications of AVSCA systems to PD and positions this work in the context of related works and literature.

4.1.1 Building blocks

An AVSCA system usually follows a pattern recognition-like structure[78]: given an input acoustic signal, characteristics are extracted in the form of a set of features to accomplish a further decision-making task downstream.

The architecture of the system is conditioned to two important design decisions[78]

- *type of input speech*, could be sustained phonation of vowels or running speech; the latter (as stated in the introduction to chapter 3) is further subdivided in spontaneous, reactive and imitative, each of these categories can be leveraged to extract different kinds of information for different kinds of diagnosis;

[1] Slightly extending the concept of Automatic Voice Condition Analysis (AVCA) proposed by [78] to cover aspects of speech not dependent only on voice or phonation

- *output decision type*, could be detection or binary classification (is distinguishing control and pathological), identification or multi-class classification (distinguishing between multiple disorders affecting speech), or assessment/grading (a score on a perceptual rating scale)

In choosing the type of input speech one may consider that sustained phonation is easier to analyze, straightforward for the speaker, allow for analysis of voice in isolation from other speech or linguistic aspects, and their simpler acoustic structure may lead to more consistent results[78].

On the other hand, using running speech as input allows for the evaluation of the performance of the articulatory system (3.1.3) and consider factors like coarticulation (3.3.1) which introduces dynamical effects that can be important for certain applications[78]. Moreover, it is desirable to analyze running speech because it is a more realistic use of voice and the speakers are less likely to compensate for voice problems[78]. Running speech enables also the analysis of suprasegmental aspects (3.3.2) of speech like variations in pitch and loudness, onsets, terminations, breaks[78].

The *input audio stream*, of a specific type of speech among the ones just discussed, usually goes through the following blocks[78]

- *preprocessing*. The audio signal is decomposed in short equal-length frames, a duration for which the process generating the signal can be assumed to be stationary; frames usually partially overlap and on each frame a window function is typically applied to improve spectral properties; voiced/unvoiced or silence detectors can be added to analyze only intervals containing phonation or actual speech respectively; filtering may be used to accentuate certain frequencies relevant to speech, but this does not necessarily improve performance of the system and can be problematic when noise-like sounds are the target of some parts of the analysis.

- *feature extraction*. Features that represent the properties of the classes under analysis are computed. The result is a vector of d features, usually based on temporal and acoustic analysis, perturbation and fluctuation, spectral-cepstral, complexity and others.

- *dimensionality reduction*. Removal of redundant or irrelevant features that might affect performance. Several methods are usable, based on singular value decomposition, linear discriminant analysis, some form of principal component analysis; a decision-making approach to feature selection can also be adopted using a performance metric obtained using a classifier/regressor, correlation and information can be used for a filter feature selection approach, while it is also possible to embed feature selection so that it is done in the training of the final model.

- *machine learning and decision-making*. Supervised or unsupervised machine learning algorithms used to evaluate the desired output. Among these, the most common include Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), Deep Neural Networks (DNNs), Hidden Markov Models (HMMs), Random Forests, Linear Discriminant analysis, k-Nearest Neighbors (k-NN), Bayesian classifier, or regression techniques.

- *evaluation of the system.* This step usually involves a test set to assess the performance of the system. Test set selection and evaluation can be performed by means of several techniques, including k-folds cross-validation (one of the more popular in AVSCA), split sample, leave-one-out, bootstrapping, cross-dataset validation; the evaluation is used to compute metrics, the simplest one used is the accuracy, while in the case of binary detection systems Receiver-Operating Curve (ROC), Detection Error Tradeoff (DET) and Area Under ROC Curve (AUC) are common.

In the end, a *decision* is obtained as the output of this flow, according to the output decision type chosen.

4.1.2 Complications

A major source of errors in AVSCA is the variability embedded in speech, examples of inter-class variability are[78]

- *dialects* (impacting phonology, morphology, syntax, lexicon and/or semantic level) and *accents* (having an impact mostly on the phonological level), with non-prestige social dialects often associated with disordered speech;
- *vocal effort* is the subjective interpretation of the level of adaptation of the speech to demands of communication, a high level of adaptation can cause large variations in the parameters examined by AVSCA, causing a normophonic voice to seem less so and vice versa;
- *emotions* which can act as a confounding factor
- *sex* and *age*, which account for most of the variability in ASR systems, can act as a confounding factor due to physiological, acoustic, and psychophysical factors.

Data concerning some of these factors can be provided as input parameters to models in the system to adjust the system behavior and support in decision-making[78].

Other influences may be external and dependent on the channel, such as differences in instrumentation, procedure, environment and transmission mean; including also the impact of noise from the environment or the speaker[78].

4.1.3 Applications to Parkinson's Disease (PD) and related work

A growing body of literature in the past decades[119, 125] has been focusing on speech and voice impairment of people with PD, mainly including automatic and computerized analysis.

In this context, several studies exist using different kinds of input speech obtained by having the patient perform different tasks, to ultimately consider mainly three aspects of voice/speech[119, 125]

- *phonatory*, related to the glottal production of sound and its resonant characteristics in the vocal tract; typically analyzed through sustained vowel sounds

- *articulatory*, related to the modifications to the speech sounds by muscles in the upper vocal tract; could be analyzed with diadochokinetic tasks, characterized by rapid, repetitive and consonant-rich speech which can highlight defective speech articulation, as well as sentence repetition and connected speech

- *communicatory*, related to the content/meaning of speech; requires a task which leads to the production of connected speech, which is usually obtained through a text-based or conversational task

In [4] Amato et al. performed a multi-level analysis progressively combining a total number of 126 features, extracted from the entire signal, voiced segments and onset/offset regions. For each of the voiced segments, features were computed on a frame-basis and on the entire signal, hence statistics were extracted to perform a dimensionality reduction. They used recordings of 25 isolated words for each subject from the PC-GITA Colombian Spanish corpus, well-balanced in terms of age and gender including 50 PD patients and 50 healthy controls, to train a k-NN model for the binary classification of PD patients. The same dataset has been used for validation and testing, obtaining a 99.4% ten-fold cross-validation accuracy and a 94.3% accuracy in testing. Note that the automatic segmentation performed in this work, aided by the Praat software, was able to distinguish voiced from unvoiced segments and provide onsets/offsets regions, but does not take into account phones or phonemes.

More recently, Amato et al. [5] analyzed the Transition Regions (TR) of specific phonetic groups to model the loss of motor control and the difficulty to start and stop movements typical of PD. They extracted 60 features from pre-processed vocal signals coming from two different datasets, Italian Parkinson's Voice and Speech (IPVS) (5.1.3) recorded in a controlled environment using professional equipment and an additional dataset recorded by patients in their homes using smartphones. The features obtained have been used to optimize two SVMs, one for the controlled environment dataset and one for the other, obtaining respectively an accuracy of $98\% \pm 1.1$ and $88\% \pm 2.8$ in 10-fold cross-validation on the respective datasets, for the task of discriminating PD patients from healthy controls. It is worth noting that the segmentation required to find the TRs in [5] is performed manually to avoid introducing a bias in the results from the intrinsic error of an automatic segmentation system[5].

The main purpose of the present work is to implement a ASR and FA system to be used in the architecture of [5] to replace the manual segmentation step, enabling further automation and thus increasing the scalability and reducing the cost of the proposed solutions. It can also add information about phonemes in the architecture of [4].

Beyond its main purpose, this work enables the analysis of spontaneous speech and different recording tasks, not requiring a transcription, and can add other useful features for classification, such as measure of speech intelligibility. Indeed this latter proved to be efficient in [50], although the intrinsic limitations caused by the use of a proprietary system developed by a third part. These limitations can be overcome with the "open" structure of the present work.

4.2 Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR), or Speech-to-Text (STT), is the task of recognizing and translating spoken language to text.

Research in ASR has attracted a lot of attention over the past six to seven decades[96]. The first ASR systems date back to 1952 when researchers from Bell Laboratories built "Audrey", a system able to recognize isolated digits from a single speaker[96]. The system, similarly to others that followed shortly after, was based on acoustic phonetics, using the formant frequencies during vowel regions of each digit. This first system was followed by some similar digit, vowel, syllable or phoneme recognizers in the 1950s and 1960s[96]. Most notably J. Sakai and S. Doshita from the University of Kyoto introduced a *segmenter* in their 1962 phoneme recognizer, enabling the use of continuous speech as input. In 1959, D. B. Fry and P. Denes incorporated statistical information about allowable phoneme sequences in English increasing the accuracy of their phoneme recognition system[96]. Later in 1964 and 1968, T. B. Martin et al. and T. K. Vintsyuk respectively, introduced a way of avoiding the use of an explicit speech segmenter in favor of the adoption of a non-uniform time scale for aligning speech patterns[96]. Both works paved the way for *dynamic time warping* and other dynamic programming based solutions, such as the Viterbi algorithm[96]. The late 1960s have seen the introduction of Linear Predictive Coding (LPC), source-filter theory (3.1.2) based model to represent or analyze speech[96]. By the mid 1970s fundamental pattern recognition techniques are applied to ASR, leveraging LPC methods[96].

During the early 1970s funding to the Speech Understanding Research (SUR) program was provided by the United States Defense Advanced Research Projects Agency (DARPA)[90, 96]. Among the systems built for this program, the Carnegie Mellon University's "Harpy" was shown to recognize speech with reasonable accuracy, using a vocabulary of 1011 words[90, 96]. In this system the input speech is put through parametric analysis and then segmented, this segmented parametric sequence is then put through template matching, a technique which had particular success in these years[96]. A graph search is then used, introduction of the beam search algorithm, on the output of the template matching to find the highest score sequence of words satisfying the lexical, syntactical and word boundary rules[90, 96].

In parallel with the efforts from SUR, IBM and AT&T Bell Laboratories carried their speech recognition research using two different approaches which became two different schools of thought on how to realize solutions with commercial viability[96].

IBM had the goal of making a "voice-activated typewriter": they developed a speaker-dependent system with a focus on the size, as large as possible, of the recognition vocabulary and the *language model*, describing how likely a sequence of words or phonemes would be[96]. Language models became fundamental in later large vocabulary ASR systems: it was shown that an n-gram model with $n \geq 3$ outperforms a human in predicting the next word in a sentence[96].

AT&T Bell Laboratories, to provide automated telecommunication services like voice dialing or phone-calls routing, focused on speaker-independent systems[96]. With speech clustering and statistical modeling techniques, the emphasis was on the *acoustic model* over the language model[96]. This led also to the development of a *keyword spotting* approach, able to recognize single words in running speech, more than sufficient for some

use cases[96].

Despite the differences, the efforts by IBM and AT&T Bell Laboratories had in common a strong mathematical formalism which, together with the profound influence they had at the time, led to the development and adoption of statistical methods in the following years[96].

4.2.1 Traditional approaches

ASR in the 1980s transitioned from template-based approaches to more rigorous statistical modeling frameworks, among those the most popular became the HMM which remained the foundation of most ASR systems until the early 2000s[96, 179].

HMMs are a formal foundation for probabilistic modeling of linear sequences "labeling problems"[55]. A Markov process is a process whose next state only depends on the current state, if a Markov process has a finite state space and discrete time-steps it is known as a Markov Chain[148]. A hidden Markov chain X has unobservable states ("hidden"). A HMM learns about this hidden chain through an observable process Y whose outcomes are influenced by X in a known way and only by the outcomes of X of the corresponding time instant[148].

A HMM could be optimized using an Expectation-Maximization algorithm like the Baum-Welch (equivalent to training in machine learning)[90, 96], evaluated using the Forward or Backward algorithms[90] and its output efficiently decoded through the dynamic programming Viterbi algorithm[55].

Applications of HMMs to ASR first used a Gaussian distribution for the observations, and later had more success, especially in speaker independent and large vocabulary performance, with the use a Gaussian Mixture Model (GMM) pioneered by Bell Laboratories[96].

In the 1990s, HMMs were extended with *finite-state grammar* to account for the structure of language at an articulation and pronunciation level, improving performance and efficiency in large vocabulary continuous speech recognition[96].

HMM-based models can generally be divided into three parts, each independent and with a different role[178]:

- the *acoustic model*, for the mapping between speech input and feature sequence (typically a phoneme or sub-phoneme sequence)
- the *pronunciation model*, typically constructed by professional human linguists, for mapping between phonemes (or sub-phonemes) to graphemes
- the *language model*, mapping the character sequence to a final transcription

These different modules usually use different technologies, HMMs are mainly used to do dynamic time warping at the frame level while GMMs are used to calculate emission probability for the HMM hidden states[178].

The 1990s have seen also the development of open research tools such as HTK, Sphinx, FSM Library and others started to be available simplifying the development of new systems[90, 96].

The 2000s have seen a shift toward ANNs. In the late 1980s, the possibility to use error back-propagation on new more powerful hardware enabled a comeback of ANNs,

especially in pattern recognition with the multi-layer perceptron able to approximate any function to an arbitrary prcision[96]. The combination of HMMs and DNNs led to significant performance improvements, with DNNs replacing GMMs and learning powerful discriminative features while leveraging all the existing research on HMMs[90]. The use of DNNs instead of GMMs led to the use of simpler features such as spectrograms and filter banks to allow the DNN to discover more useful representations on its own[90].

Moreover, the massive amount of data starting to become available enabled improvements in acoustic and language models[90]. For example, eliminating the need for explicit segmentation and labeling of phonetic strings[90].

As DNNs were replacing GMMs improving acoustic models, Recurrent Neural Networks (RNNs) were a significant improvement over N-gram language models and variants[90].

HMM-based models have been the state of the art for ASR until relatively recently, but that was not without difficulties in their practical use[178]

- *complex training process difficult to globally optimize*, due to the different training methods and datasets used by the different parts of the model; each module is optimized individually not necessarily converging toward the global optimality for the ASR system
- *conditional independent assumptions* are used to simplify the model and its training, but they are often not realistic in ASR systems

4.2.2 End-to-End (E2E) approaches

The wide availability of large quantities of data and the successful adoption of deep learning technologies for other tasks, coupled with the shortcomings of HMM-based and traditional models led to more and more works studying End-to-End (E2E) systems[178].

This new type of system and the corresponding model are called end-to-end because of their two main characteristics and related advantages[105, 178]

- they *directly map* an input audio sequence to an output sequence of words or symbols (they do not require further processing to achieve the true transcription or to improve recognition performance)
- they *merge multiple modules* into a single deep network
  - enabling *joint training*, allowing the use of a global optimization goal that is more relevant to the final evaluation criteria
  - removing the need for many modules and the mapping between carefully-designed intermediate states

Even though the model is a single deep network it is usually decomposed into different parts with different subtasks (note that there is no clear division between them in the network as there would be in traditional models) these include[178]

- an *encoder*, realizing the mapping of the input speech sequence to a feature sequence;

- an *aligner*, which does the alignment between the feature sequence and language;
- a *decoder*, which decodes the output of the encoder and the aligner into the final output of the model

Another advantage of E2E models is that they do not require an explicit alignment of the audio to the text, using a *soft alignment* approach that associates each audio frame to all possible labels with a certain probability distribution[178]. E2E models can be categorized based on their implementation of soft alignment[105, 178]

- *Connectionist Temporal Classification (CTC)-based*, which enumerates and then aggregates all possible hard alignments while assuming that the output labels are independent of each other, this last assumption implies that a CTC cannot incorporate a language model, but studies have obtained good results integrating it on top
- *RNN-transducer*, enumerates and then aggregates all possible hard alignments but does not assume independence among output labels and thus can learn the language model
- *Attention-based*, uses the Attention mechanics to directly compute the soft alignment information between input audio and output label, also in this case the model can learn the language model

Hybrid architectures are possible too, for example, [98] uses a combination of CTC-based and Attention-based alignment.

4.2.3 Supervised, semi-supervised, self-supervised and unsupervised approaches

The methodologies for ASR described in 4.2.1 and even those described in 4.2.2 may have relaxed the requirements on aligned labels for training but still require transcribed speech, which makes them unusable for the vast majority of languages and limits their performance for others[13]. This is because they have been typically trained with a *supervised* approach, meaning that for each audio recording in the dataset the model is provided the corresponding transcription during training; however, datasets with these characteristics are relatively expensive to realize and are not available for the majority of languages, or they are not large enough to obtain good results with simple supervised approaches[13].

To leverage data with no transcription associated, more widely available for more languages, three main approaches have emerged[13]

- *semi-supervised* ASR through *self-training*, which consists in using *pseudo labels*, the output of a ASR model previously trained with a supervised approach, for each recording and then proceed with training as in supervised approach
- *self-supervised*, defining and using *pretext* tasks to learn useful representations of speech audio from audio only

- *unsupervised*, learn from raw audio signal only to identify meaningful units in the sound stream[2]

Starting from 2015, the Zero Resource Speech Challenge (ZRSC) has the goal of accelerating research in the field of unsupervised speech processing[2]. The interest in unsupervised approaches is also motivated by the fact that humans learn a lot about speech just by listening to others around them talking without explicit supervision, even if other environmental and interactive factors certainly play a role[2, 13].

With no labels, unsupervised learning can only discover recurring patterns in the input audio signal and learn the relationships between those patterns[2]. The model learns abstract representation that need other learning approaches to be adapted to orthographic or phonetic form[2].

4.2.4 State of the art

At the time this thesis work began, to the best of the author's knowledge, the XLS-R model presented in [12] was the state of the art for ASR on Italian.

Later, during the development of this work, [53] found that UniSpeech-ML[177] is slightly better than XLS-R which however is found to be the second-best model for ASR on Italian and stands very close to UniSpeech-ML in terms of Word Error Rate (WER) (5.2.2) and Character Error Rate (CER) (a deeper and more technical discussion on XLS-R can be found in 5.3.1).

Shortly after the publication of [53], OpenAI released Whisper[149] which is pretrained on even more data than XLS-R and outperforms it and Wav2Vec 2.0[14] in some benchmarks.

4.2.5 Non-normophonic speech

The variability in speech due to the speaker dialect, provenience or social context is usually a complication or confounding factor for ASR, due to under-representation in the datasets on which they are trained.

Non-normophonic speech is even more problematic for common general purpose ASR products or services, because the large datasets typically used to develop and maintain these systems contain recordings from people with unimpaired speech[45, 160].

The impact of dysarthric English speech from the TORGO database[155] on popular commercial ASR systems was recently evaluated by De Russis and Corno [45], which found[45]

- for Google Cloud Speech, a WER of 16.11% for dysarthric users with "no abnormalities" and 78.21% for "severely distorted", while a WER of 3.95% is obtained on a healthy control group;
- for Microsoft Azure Bing Speech, 23.16% and 78.59%, 6.94% on control;
- for IBM Watson STT, 14.89% and 89.08%, 5.26% on control.

A similar study has been done by Ballati, Corno, and De Russis [16] on the impact of Italian dysarthric speech on the performance of voice assistants, finding an average

WER of 24.88% for Google Assistant, 39.39% for Microsoft Cortana and 70.89% for Apple Siri.

Oddly enough, Jefferson [93] found that this poor performance of ASR systems on individuals with speech disorders may actually be a regression, with pre-2011 systems seemingly performing better than current ones.

Recently there have been some studies and efforts aimed at improving ASR for individuals with dysarthric speech or other speech disorders[76, 87, 94, 109, 145, 150, 162, 166, 168].

In 2019[162], Google Research announced an initiative focused on helping people with non-standard speech be better understood, called Project Euphonia, and in this context has then collected over 1300 hours from more than 1000 individuals with speech disorders[109, 145].

Shortly after, Shor et al. [162] used a RNN-Transducer, in a configuration with a bidirectional encoder without attention, trained on the Librispeech[132] English corpus and then fine-tuned to obtain a personalized model for a single speaker, using the data they just started collecting in Project Euphonia. The fine-tuning resulted in an improvement over the base model from 59.7% to 20.9% WER for more severe speech disturbance, and from 33.1% to 10.8% for less severe ones[162]. Another interesting finding is that 71% of the improvement can be obtained with only 5 minutes of data for fine-tuning[162].

Later Green et al. [76] refined the approach from [162], in particular introducing the use of layer freezing and SpecAugment[133] in the personalization/fine-tuning phase. The accuracies of personalized models were significantly better than those of the speaker independent models used for comparison, 4.6% median WER against 31% median WER, personalized models outperformed even human listeners, especially for moderate to severe conditions[76].

In a setup similar to [76], Tobin and Tomanek [166] focused on using small amounts of per-speaker adaptation data. They fine-tuned the personalized models with small increments of speech data until a target WER was obtained. This approach, in an home automation scenario, required only 3 to 4 minutes of speech to reach the target WER for 63% of speakers, 18-20 minutes for 79%[166].

These last 3 works have all been developed using data from Project Euphonia, which at the moment is not available to the public and covers mainly English, with some preliminary steps on French, Hindi, Japanese, and Spanish[146].

Having access to small datasets, it can be useful to leverage data augmentation techniques (e.g. recently Jin et al. [94] introduced a new technique based on deep convolutional generative adversarial networks to modify normal speech spectra to resemble those obtained from disordered speech).

4.3 Forced Alignment (FA)

Forced Alignment (FA) is the task that, given a speech signal and its orthographic transcription, automatically aligns speech and the transcription at the word and/or phone level, usually requiring a way to map graphemes to phonemes and a statistical model of how phones are realized[20, 114]. It can be decomposed in the sub-tasks of[20]

- *phonetization*, also known as Grapheme-to-Phoneme (G2P), the process of representing text by phonetic signs
- *alignment*, the process of aligning speech with these signs

Several forced aligners have been developed during the past few decades[21, 70, 74, 114, 124, 151]. Due to the availability of these software solutions and their relative ease of use, forced alignment has become widely used in scientific research on language over the past 15 years, including the fields of sociolinguistics, phonetics, language documentation, and psycholinguistics[114].

4.3.1 Methodologies and state of the art

Differentiating the available solutions there are[137]

- architecture and the underlying algorithms and models, traditional solutions are mostly based on HMMs while recent works are exploring the use of E2E models
- languages supported, most of them support English, some of them support other languages or support being trained/ported on other languages
- trainability, most of the solutions available have pre-trained models which are not modifiable, some support training models with new data, from scratch or starting from pre-trained models
- license, ranging from proprietary/closed source to more or less permissive open source licenses

Historically, as in ASR systems, HMMs have been used at the core of FA systems[21, 70, 74, 114, 124, 142, 151], mostly with GMMs.

At the beginning of this thesis work, the Montreal Forced Aligner (MFA) as presented in [114] and later updated[117], was regarded as the state of the art in FA for English[101, 104].

MFA uses the Kaldi ASR toolkit[142], a framework similar to the Hidden Markov Model Toolkit (HTK)[187] but with a more permissive license allowing for easier distribution of executables removing the need for source compilation[114]. MFA adapts a standard GMM-HMM ASR pipeline from Kaldi: it trains and uses monophone GMMs to obtain a first alignment; then uses these alignments to train triphones GMMs that take into account the surrounding phonetic context. The alignments generated from the triphones GMMs can then be used to train a model with speaker adaptation[114]. MFA derives from the Prosodylab-aligner and maintains one of its key features: its trainability, which allows for it to be trained to work on languages different from English[114]. Its evaluation on the Buckeye[138] English corpus results in a phone level boundary accuracy of 77% and 93% for a tolerance of 25ms and 50ms respectively, for word level boundaries an accuracy of 68% and 88% respectively for a tolerance of 25ms and 50ms is reported, see table 1 of [114] for more details.

As in ASR, E2E models introduce architectural improvements, they are deeper, and they have many more parameters, adding model capacity[98].

However, higher model capacity and parameters count require larger training datasets to achieve comparable levels of performance[98].

As we have seen for ASR, E2E models solve this requirement partially thanks to the availability of larger datasets compared with the past and mostly leveraging unlabeled data using self-supervised pre-training or semi-supervised training on pseudo-labels.

Kürzinger et al. [98] uses a pre-trained CTC-Attention hybrid ASR model from ESPnet[180] to pseudo-label more training data from unsegmented or unlabeled datasets. They introduce and use *CTC-segmentation* "an algorithm to extract proper audio-text alignments in the presence of additional unknown speech sections at the beginning or end of the audio recording"[98], the same algorithm used in this work. The CTC-segmentation can be used on top of a CTC-based ASR model[98]. They evaluated its performance, using the pre-trained ASR model previously mentioned, on the sentence-level manually aligned TEDlium v2[152] corpus and compared it against HMM-based Munich AUtomatic Segmentation (MAUS)[124] and Gentle[70] and Dynamic Time Warping (DTW)-based Aeneas[1, 98]. Kürzinger et al. obtained significant improvements against MAUS and Aeneas in terms of mean deviation from the ground truth, 0.31-0.35s against 1.38s and 9.01s, and a percentage of boundaries with an error smaller than 0.5s, 85.1-90.1% against 74.1% and 64.7%[98]. The accuracy of MAUS and Aeneas is further reduced when additional unknown parts are added at the beginning and end of each sample[98]. Gentle[70] is closer to the performance of the system from Kürzinger et al., but it still has a worse mean deviation at 0.41s and lower 0.5s-tolerance accuracy at 82.0%[98]. Differently from MAUS and Aeneas, it was more robust to the addition of unknown audio parts[98].

Hira [84] included an implementation of the algorithm described by Kürzinger et al. [98] and related documentation in PyTorch[136] Torchaudio[186] library, enabling easier adoption of this methodology of FA.

More recently Li et al. [104] used a multi-task learning approach for their "NeuFA" E2E model based on bidirectional attention: they pre-trainined the model on the Librispeech[132] English corpus and then had the model learning the bidirectional relationship between text and speech solving simultaneously the two separate but related tasks of Text To Speech (TTS) and ASR. They evaluated their model on a subset of the Buckeye[138] English corpus against MFA[114], having both of them trained on another portion of Buckeye[104]. NeuFA is found to have a 23.7ms Mean Absolute Error (MAE) at word level and a 15.7 one at phoneme level, a slight improvement against the 25.8ms and 18.0ms respectively obtained from MFA[104]. Their accuracies for word level and phoneme level alignments at 10ms, 25ms, and 50ms tolerances are also improved respectively by 14%, 4%, 1% and 5%, 3%, 1%[104].

Very recently López and Luque [107] developed a system based on CTC-segmentation[98] and used it to align the RTVE2022DB[42] Spanish database. They do no report metrics on the quality of the alignment.

Another interesting development in FA has been into its cross-language use, mainly to make reasonable quality FA available also for languages with less or no transcribed data[101]. This task is called Cross-Language Forced Alignment (CLFA)[101]. MAUS has a language-independent version[101] and MFA has added a "multilingual IPA" mode[116].

4.3.2 Forced Alignment (FA) on Italian speech

Concerning the use of FA specifically on the Italian language, the earliest article that the author could find is [32].

Even in the early 1990s, in laboratory conditions, automatic segmentation of Italian speech had reached very good performance with Angelini et al. [6] obtaining, with a 20ms tolerance, a 90.9% accuracy on a subset of the APASCI[7] Italian corpus and an 86.2% accuracy on a subset of the TIMIT[69] American English corpus[26]. They used an acoustic-phonetic unit recognizer based on a HMM combined with a rule-based network to cope with pronunciation variability, finally, the Viterbi[173] algorithm has been used to determine the most likely phone sequence and phone boundaries[6].

In January 2011, Cangemi et al. [26] reported a 94% accuracy on a larger subset of APASCI with a 20ms tolerance and a 99% one with 30ms tolerance. They leverage the Speech Processing, Recognition and Automatic Annotation Kit (SPRAAK)[47] open-source ASR package to train a GMM-HMM-based acoustic model on a portion of the Corpora e Lessici di Italiano Parlato e Scritto (CLIPS) corpus (described in detail in 5.1.2) combined with Vocal Tract Length Normalization (VTLN), Cepstral Mean Subtraction (CMS), and Mutual Information Discriminant Analysis (MIDA) to increase robustness for speaker and channel variability[26].

Shortly after, the 2011 edition of EVALITA[2], "a periodic evaluation campaign of Natural Language Processing (NLP) and speech tools for the Italian language"[59], included a FA task[38].

The EVALITA 2011 FA task provided as training and validation datasets an adapted version of the dialogic subcorpus of the freely available CLIPS corpus and used as test-set an unreleased batch[3] consisting of approximately 10 minutes of dialogic speech recorded and labeled during the development of CLIPS[38, 106]. The task is further divided into two sub-tasks, according to the segmentation level[38]

- *word segmentation* and
- *phone segmentation*,

both available in two modalities[38]

- *closed*, allowing training and tuning only on the material provided by the organizers, and
- *open*, allowing the use of additional material.

There were three teams participating in this task[38]: Bigi [22], Ludusan [106] and Paci, Sommavilla, and Cosi [130]. Table 4.1 provides an overview of the results for the word segmentation tasks, both in open and closed modalities. Table 4.2 provides an overview of the results for the phone segmentation task, both in open and closed modalities.

[2] https://www.evalita.it/

[3] The author reached out to the organizers of the EVALITA 2011 FA task to obtain a copy of the test-set, but unfortunately, according to them, the data that was used as test-set was not preserved.

According to Cutugno, Origlia, and Seppi [38, 40], the evaluation was conducted using the time-mediated alignment functionality provided by the Score Lite (SCLITE) tool in the National Institute of Standards and Technology (NIST) Scoring Toolkit (SCTK)[161]. Time-mediated alignment replaces the default weights used for scoring by the dynamic programming algorithm used by SCLITE with[40, 161]

$$D(correct) = |T_1(ref) - T_1(hyp)| + |T_2(ref) - T_2(hyp)|$$

$$D(insertion) = T_2(hyp) - T_1(hyp)$$

$$D(deletion) = T_2(ref) - T_1(ref)$$

$$D(substitution) = |T_1(ref) - T_1(hyp)| + |T_2(ref) - T_2(hyp)| + 0.001$$

where $T_1(x)$ is the beginning time mark of word x and $T_2(x)$ is the ending time mark of word x[40, 161]. The distance D for the insertion or deletion of the NULL token '@' is set to 0.001[40, 161]. Moreover, for some circumstances alternate transcriptions are provided, that is the scoring system has been configured to accept as correct slightly different transcriptions or alignments for the same reference[38]. For example, it is not considered an error to insert a /tS/ instead of a /t/tS/, or in the case of groups of three neighboring vowels, all possible groups of the three symbols are allowed[38].

The best results, according to the data from [38] reported in tables 4.1 and 4.2, were those obtained by Ludusan [106].

The system developed by Ludusan [106] is also built on SPRAAK, it uses 43 context-independent phone models based on Semi-Continuous HMM (SCHMM), in which HMM states are globally tied, with GMMs[106]. MIDA and the Viterbi algorithm are used during training[106]. In table 3 of [106], Ludusan reports some statistics on the shift of their system output boundary marker from the reference[4], shown here in table 4.3.

Participant	Mode	Corr. %	Subs. %	Dele. %	Inse. %	Error %	Sentences with errors %
Bigi	Closed	97.6	1.0	1.4	1.4	3.8	17.8
Ludusan (5ms)	**Closed**	**99.3**	**0.1**	**0.5**	**0.5**	**1.2**	**6.7**
Ludusan (10ms)	Closed	99.2	0.2	0.6	0.6	1.4	7.8
Ludusan (5ms)	Open	99.0	0.2	0.8	0.8	1.8	10.0
Ludusan (10ms+VTLN)	**Open**	**99.3**	**0.2**	**0.5**	**0.5**	**1.2**	**5.6**
Paci	Closed	98.4	0.4	1.2	1.2	2.8	16.7
Paci	Open	97.4	1.2	1.5	1.5	4.1	14.4

Table 4.1. EVALITA 2011 FA task results of word segmentation, from tables 1 and 5 of [38]

[4]The author could not find an indication on which data (CLIPS portion or training set, unpublished CLIPS portion or test set, other data) was used to compute these statistics

Participant	Mode	Corr. %	Subs. %	Dele. %	Inse. %	Error %	Sentences with errors %
Bigi	Closed	83.7	11.3	5.0	4.9	21.2	93.9
Ludusan (5ms)	Closed	93.0	5.0	2.0	8.1	15.1	80.5
Ludusan (10ms)	**Closed**	**93.9**	**4.9**	**1.2**	**7.2**	**13.3**	**79.8**
Ludusan (5ms)	Open	93.0	5.2	1.8	8.2	15.1	81.6
Ludusan (10ms+VTLN)	**Open**	**93.6**	**5.1**	**1.3**	**7.2**	**13.6**	**79.1**
Paci	Closed	92.4	5.9	1.7	4.5	12.1	81.0
Paci	Open	90.6	7.3	2.1	4.6	13.9	81.3

Table 4.2. EVALITA 2011 FA task results of phone segmentation, from tables 3 and 6 of [38]

Level	$\leq$ 10ms [%]	$\leq$ 20ms [%]	$\leq$ 30ms [%]	$\leq$ 40ms [%]	$>$ 40ms [%]
Phone	56.59	82.20	93.78	96.86	100
Word	46.04	70.46	86.03	91.02	100

Table 4.3. Ludusan [106] results for EVALITA 2011 FA task, from table 3 of [106]

All the participants obtained results close to the state of the art for other languages even if the CLIPS corpus was reported to be particularly challenging[38].

The results from Ludusan in table 4.3 are seemingly worse than those from Cangemi et al. [26] using the same 20ms tolerance (82.2% vs 94%), but it must be taken into account that the dialogic sub-corpus of CLIPS, being composed of spontaneous speech, is much more challenging than APASCI. Moreover, the EVALITA 2011 FA did not provide a phonetic transcription to align but just the corresponding graphemes, introducing the further challenge that is the development of a system that considers several pronunciation variants in the G2P sub-task.

The 2014 edition of EVALITA has seen the general FA task replaced by a task on FA of children's speech specifically[60, 67]. This FA task has seen only the participation of the SPPAS team[21, 36], and was structured similarly to the 2011 one but used the CHILDIT-2[34] corpus. The transcriptions for the CHILDIT-2 corpus, differently from the CLIPS corpus, are not manually generated; they are automatically obtained using a, then state-of-the-art, a Kaldi-based GMM-DNN ASR system trained by [35] on the CHILDIT[33] corpus[67]. No evaluation or test results have been found on the FA capabilities of the system from [35]. The results of the SPPAS team for this challenge[67] are similar or slightly better than the ones obtained by [106] in the EVALITA 2011 FA task, but given the differences in the corpora used the comparison is not really fair.

To the best of the author's knowledge, no work describing E2E FA solutions for Italian has been published yet.

4.3.3 Complications

FA is a difficult task. The accuracy of automatic FA systems is normally evaluated on manually labeled and segmented speech signals from phonetic or speech communication experts, but even these golden reference labels are implicitly incoherent due to variability of visual and acoustic perceptual capabilities and difficulties in defining and using a clear common labelling strategy[32]. Moreover, this kind of labeling and segmentation is time-consuming and expensive, thus very few corpora exist of this kind that are public or free to use. On top of that, these resources are typically available only for a small number of the thousands of languages spoken around the world[20].

The FA task usually takes in input a speech signal and its transcription, a sequence of graphemes that in this case should be converted to a sequence of phones or phonemes. The G2P sub-task doing this conversion should take into account all the different ways in which the same utterance can be pronounced, due to different accents or different speaking rates[20]. Speech corpora use conventions to establish which phenomena to transcribe and how[20].

G2P conversion is even more challenging on spontaneous continuous speech, which is characterized by an important gap in words' phonological form and their actual phonetic realizations[20]. Spontaneous speech presents frequently elision or reduction, but also substitutions or addition of phonemes, which have an impact on the automatic phonetization and alignment sub-tasks, in some languages code-switching can add to the complications too[20].

G2P is historically performed with a dictionary-based, rule-based or some intermediate approach[20]. In the EVALITA 2011 FA task, the G2P sub-task has been identified as the one needing refinement if not deep review to work on spontaneous speech data, like the dialogic sub-corpus of CLIPS[22, 38].

Another difficulty in the practical application of FA systems is that they typically assume that the audio contains only the text which should be aligned, but for several use cases and most public domain data this is not the case[98]. On longer audio tracks, differences from the transcription can easily lead to a largely misaligned result due to incapacity to recover[107]. Newer E2E FA systems can however mitigate or solve this specific problem[98, 104].

Chapter 5

Overall methodology and materials

This chapter provides a brief general description for each corpus, dataset, metric and model used in this work; it introduces their role in this work and then provides an overview on the architecture and the methods used for the implementation.

5.1 Corpora and datasets

The different tasks composing this work have imposed some common and some specific requirements on the datasets to be used:

- all the datasets must be in Italian, or have a large enough Italian portion
- at least one dataset must provide accurate transcription of all its audio data
- at least one dataset must provide human annotated phoneme-wise alignment labels
- at least one dataset must include disordered speech caused from PD with relevant metadata

In the end three corpora/datasets have been singled out: CommonVoice (5.1.1), CLIPS (5.1.2) and IPVS (5.1.3).

5.1.1 Common Voice

The Common Voice project was started by Mozilla in July 2017[8, 122], the goal of the project is to provide a free and open source corpus to make speech technology research and development more accessible, with particular attention to ASR but enabling also other kinds of developments (e.g. language identification)[8]. The goal is being achieved in a sustainable and scalable way by employing crowdsourcing for both data collection and data validation[8].

Contributors can participate using either the Common Voice website or smartphone app, recording their voice while reading sentences displayed on the screen or validating recordings from other speakers[8].

The corpus is thus composed of reading speech, the text read comes from Wikipedia articles, for languages with more than 500000 articles, or is contributed by users; in both cases, an automatic system splits the text into sentences and filters them using a set of rules, after that at least 2 out of 3 reviewers must approve them[8]. If approved they are served as prompts to be read for contributors.

Recordings from contributors are then reviewed by other contributors. A maximum of three contributors will listen to any audio clip. If the audio and transcript pair receives two upvotes, then the clip is marked as valid, if instead the clip first receives two downvotes then it is marked as invalid[8].

The first release of the Common Voice corpus dates back to November 2017, containing about 500 hours of English speech from 20000 people[83, 184]. Since then many releases followed adding more hours of speech and more languages. The project now aims to release a new version every 3 months, at the time of writing the latest release is Common Voice Corpus 10.0[122].

Released the July 4th 2022, version 10.0 contains 20,817 hours of recorded speech, of which 15,234 have also been validated, in 96 different languages[122]. Among these languages, we have[123]

- English: 2,275 hours validated of 3,050 total hours recorded, from 83,790 unique speakers
- Italian: 321 hours validated of 347 hours recorded, from 6,735 unique speakers

Each entry in the dataset consists of a unique MPEG-3 file with a 48kHz sampling rate and a row in a Tab-Separated Values (TSV) file containing the corresponding text and other parameters[8, 122]. Indeed, many of the 20,817 recorded hours in the dataset also include demographic metadata like age, sex, and accent that can help train the accuracy of speech recognition engines[122]. The corpus is divided into train, test, and development sets bucketed such that any given speaker may appear in only one and saved as separate TSV files[8].

Since the first release, it has been used, often in conjunction with other corpora or datasets, to train and evaluate several ASR systems. Some examples are DeepSpeech[82] and XLS-R[12].

In this work, the Common Voice corpus is used for the fine-tuning of a pre-trained XLS-R model and for part of its evaluation for the ASR task, section 5.4 and part II deal with this activities in more details.

5.1.2 Corpora e Lessici di Italiano Parlato e Scritto (CLIPS)

The Corpora e Lessici di Italiano Parlato e Scritto (CLIPS) corpus, in English "Corpora and Lexicons of Spoken and Written Italian", has been developed from 1999 to 2006 as part of a project funded by the Ministero dell'Università e della Ricerca Scientifica e Tecnologica (MURST), the Italian Ministry of University and Scientific and Technological Research (now Ministero dell'università e della ricerca (MUR))[39, 102].

The project was aimed at the development of tools for the general study and automatic treatment of the Italian language in written or spoken form[102]. The CLIPS corpus focuses on the spoken dimension of the Italian language to enable[39, 102, 158]

- the analysis and description of the spoken language in all the conditions it is used
- the development of tools and applications to build automatic systems for speech recognition and production

For these purposes, the corpus is composed not only of recordings but also transcriptions and *labels* for a portion of the recordings. Transcriptions contain the text (as graphemes) of the speech in a recording, while labels consist of the list of time-aligned phones and/or phonemes pronounced in a recording.

According to [159] the corpus contains 101 hours 36 minutes and 51 seconds of audio recordings in total, divided into 5 sub-corpora per type of speech, a portion of each sub-corpus is transcribed and/or labeled as described in[39, 102, 159]

- *Dialogico* (dialogic), contains elicited speech between couples of speakers obtained through the "map task" or "spot the difference"
  - about 30% of the collected material is transcribed
  - about 30% of the transcribed material is labeled (about 9% with respect to collected material)
- *Radiotelevisivo* or *RTV* (from radio and television), contains speech from radio or television programs like news, interviews and talk shows
  - about 30% of the collected material is transcribed
  - about 30% of the transcribed material is labeled (about 9% with respect to collected material)
- *Letto* (read-aloud speech), single words and sentences, from several speakers
  - about 30% of the collected material is transcribed
  - about 30% of the transcribed material is labeled (about 9% with respect to collected material)
- *Ortofonico* (orthophonic[1]), read-aloud speech from trained speakers, to be used as a gold reference of correct spoken Italian
  - 100% of the collected material is transcribed
  - about 16% of the collected material is labeled
- *Telefonico* (speech over a telephone line), sampled at 8000 Hz
  - 100% of the collected material is transcribed

[1]From orthophony: the art of correct articulation/speech; voice training[128, 129]

– about 3.5% of the collected material is labeled

In tables 5.1 and 5.2 are illustrated the durations by locality and sub-corpus according to [159], respectivelly overall durations and durations for labeled recordings. In table 5.3 total duration and percentages for transcribed and labeled parts of each sub-corpus as reported by [39]. In section 6.3.1 however, the use of a custom-made script to compute the values illustrated in tables 5.1, 5.2 and 5.3, directly from the corpus files just downloaded, leads to results that slightly differ from the ones presented by [159] and [39].

Locality	Dialogico	RTV	Letto	Ortofonico	Telefonico
Bari	2:18:58	0:49:44	0:51:12	/	1:12:30
Bergamo	3:16:00	0:56:09	0:44:56	/	0:59:00
Cagliari	3:13:00	1:02:42	1:39:00	/	1:16:19
Catanzaro	2:34:00	0:51:00	0:56:13	/	0:50:07
Firenze	4:42:00	0:47:19	1:25:49	/	1:15:16
Genova	2:20:00	0:49:51	0:52:07	/	1:00:51
Lecce	2:08:00	0:57:55	0:52:03	/	1:04:48
Milano	3:15:00	0:52:46	1:18:57	/	1:14:40
Napoli	2:54:00	0:54:01	0:49:44	/	0:55:01
Palermo	3:11:00	0:49:17	0:58:34	/	1:14:37
Parma	3:31:00	0:55:22	1:23:09	/	1:01:49
Perugia	3:25:00	0:56:07	0:58:36	/	1:09:41
Roma	3:55:00	0:56:16	0:51:04	/	1:17:48
Torino	4:06:00	0:57:05	1:18:22	/	1:05:55
Venezia	3:25:00	0:46:08	1:21:02	/	1:03:30
National	/	3:16:00	/	3:42:28	/
Total	**48:13:59**	**16:37:42**	**16:20:51**	**3:42:28**	**16:41:52**

Table 5.1. CLIPS corpus overall recordings total duration per locality and category as reported in attachment 6 of [159], expressed as hours:minutes:seconds

The corpus has been developed taking into account the strong variability of natural languages in their uses, with specific attention to four variables[102]

- *region*, speech changes based on the region the speaker comes from
- *social context*, speech changes based on the level of education, job, and social environment of the speaker
- *style*, speech changes depending on the situation, for example, the professional speech of a television host differs from the informal and spontaneous speech of a conversation with a friend
- *individual*, speech changes based on anatomic characteristics and idiosyncrasies[2]

[2]In linguistics the term can be applied to symbols, words, or structures, indicating them as peculiar or particular to an individual; examples of these are words or structures invented by the speaker

Locality	Dialogico	RTV	Letto
Bari	00:17:15.413	00:04:06.625	00:05:59.700
Bergamo	00:23:29.046	00:05:28.632	00:05:30.906
Cagliari	00:21:38.815	00:04:40.345	00:07:04.304
Catanzaro	00:18:03.248	00:05:06.306	00:07:30.316
Firenze	00:26:41.000	00:04:10.025	00:08:54.248
Genova	00:15:54.515	00:04:05.605	00:05:47.624
Lecce	00:20:18.561	00:04:02.268	00:09:53.795
Milano	00:28:41.000	00:04:07.669	00:11:14.586
Napoli	00:19:59.578	00:04:09.295	00:06:13.992
Palermo	00:17:34.796	00:16:50.609	00:07:02.458
Parma	00:23:13.000	00:04:22.704	00:11:26.004
Perugia	00:24:14.609	00:04:25.906	00:08:29.880
Roma	00:25:29.137	00:04:49.000	00:05:26.067
Torino	00:23:42.000	00:04:15.335	00:10:07.813
Venezia	00:19:56.000	00:04:22.768	00:09:27.549
National	/	00:02:51.958	/
Total	**05:26:10.718**	**01:21:55.050**	**02:00:09.242**

Table 5.2. CLIPS corpus labeled recordings total duration per locality and category calculated from attachment 8 of [159], expressed as hours:minutes:seconds.milliseconds

		Dialogic	Read speech	RTV	Telephonic	Orthophonic
Transcription	%	30%	30%	30%	100%	100%
	Time	15h 30’	5h 20’	4h 30’	16h 40’	3h 40’
Labelling	%	10%	10%	10%	3.5%	16%
	Time	5h 30’	1h 40’	1h 10’	35’	35’

Table 5.3. CLIPS corpus percentage and duration of transcribed and labeled material per sub-corpus from [39]

of single speakers

For this reason, the corpus presents, in appropriate proportions, recordings to reflect variations coming from these variables. According to [39, 102, 163]

- the variations due to region and social context have been taken into account in the choice of the locations where speakers have been recruited
  - the study described in [163] lead to the selection of 15 cities: Bari, Bergamo, Cagliari, Catanzaro, Firenze, Genova, Lecce, Milano, Napoli, Palermo, Parma, Perugia, Roma, Torino, and Venezia.
- the variations due to style are accounted for by the collection of different speech

tasks, read-aloud speech or spontaneous dialogic speech, and the inclusion of professional speech from radio and television

- at the individual level the dataset has a balanced number of male and female speakers, but it lacks adequate variability in the age of the speakers (in some tasks ranging only from 18 to 30 years old)

A characteristic of the CLIPS corpus that differentiates it from other public corpora of Italian speech, in the context of this work, is the presence of manual annotations, following standards common to other similar corpora also in other languages, including time-aligned segmental labeling performed by experienced operators[39].

The following paragraphs provide a brief description for the transcriptions and time-aligned labels provided by CLIPS, the labels associated with the audio file `DGmtA01T_p1F%23189.wav` are presented as examples.

Segmental time-aligned labeling includes the following levels (from narrower to broader), in different files with a specific extensions but the same file name as the ".WAV" or ".RAW" file they reference[39, 158]

- acoustic (".ACS" files), sub-phonemic level, implemented only for occlusive and affricate consonants to signal with markers the start of silence or the end of silence or of the release phase, to distinguish them from the intentional or physiological silent breaks

DGmtA01T_p1F%23189.acs

```
0 26158 __
26158 27177 d_cl
27177 27630 d_rl
27630 30621 __
30621 31183 k_cl
31183 31950 k_rl
31950 42679 __
42679 43750 k_cl
43750 44341 k_rl
44341 47284 __
47284 47678 t_cl
47678 47806 t_rl
47806 50655 __
50655 51635 tts_cl
51635 54120 tts_rl
54120 72113 __
72113 73190 t_cl
73190 73679 t_rl
```

- phonetic (".PHN" files), coded in X-SAMPA with some adaptations, it is a relatively broad time-aligned phonetic transcription, but it also represents in a narrower form a closed set of specific phonetic phenomena

DGmtA01T_p1F%23189.phn

```
0 570 __
570 2753 s
```

```
2753 5792 i
5792 15524 __
15524 26158 __
26158 27630 d
27630 29814 u
29814 30621 N
30621 31950 k
31950 33469 w-eE
33469 35985 f
35985 37857 wO
37857 38596 r
38596 39356 !i
39356 40210 D
40210 40970 a
40970 41682 &l
41682 42679 a
42679 44341 k
44341 45290 O
45290 47284 s
47284 47806 t
47806 48138 !e
48138 49610 ll
49610 50655 a
50655 54120 tts
54120 56826 jO
56826 57633 n
57633 58583 !e
58583 60814 s
60814 61384 u
61384 61763 &l
61763 62713 a
62713 65229 s
65229 66558 ua
66558 67603 D
67603 70878 E
70878 72113 s
72113 73679 t
73679 74747 r
74747 75958 a
```

- phonological or citation forms (".STD" files), coded in Speech Assessment Methods Phonetic Alphabet (SAMPA) for Italian

DGmtA01T_p1F%23189.std

```
0 570 __
570 5792 s"i
5792 15524 <sp>
15524 26158 <eeh>
26158 32492 d"unkwe%
32492 33469 %"E
33469 39356 fw"Ori
```

```
39356 42679 d"alla
42679 58583 kostellattsj"one
58583 62713 s"ulla
62713 66558 s"ua
66558 75958 d"Estra
```

- lexical or orthographic labeling (".WRD" files), includes also labels for breaks, noises, disfluencies and similar

DGmtA01T_p1F%23189.wrd

```
0 570 __
570 5792 sì
5792 15524 <sp>
15524 26158 <eeh>
26158 32492 dunque%
32492 33469 %è
33469 39356 fuori
39356 42679 dalla
42679 58583 costellazione
58583 62713 sulla
62713 66558 sua
66558 75958 destra
```

- additional level or extra-text (".ADD" files), contains extra-lexical information, such as comments, overlaps, kind of voice (e.g. creaking or screaming)

The authors of CLIPS report that labeling was done in order, starting from phonetic and acoustic, moving then to phonological and lexical, and finishing with the extra-text[158]. The theory and standards adopted by the authors did not require an alignment of the markers among the different levels, so it may happen that the boundaries for a word at the phonological or lexical level do not match the boundaries of their first and last phones at the phonetic level[158].

For each audio file belonging to the transcribed part, CLIPS also provides the following non-time-aligned labels, distributed in files with different extensions but the same file name as the ".WAV" or ".RAW" file they reference[158]

- base for orthographic labeling (".WR_" files), used as starting point for the making of orthographic labeling (".WRD" files)

DGmtA01T_p1F%23189.wr_

```
sì <sp> <eeh> dunque è fuori dalla costellazione sulla sua destra
```

- base for phonological or citation forms (.ST_ files), obtained automatically using an algorithm for grapheme-to-phoneme conversion staring from orthographic labeling (".WR_" files)

DGmtA01T_p1F%23189.st_

```
s"i <sp> <eeh> d"unkwe "E fw"Ori d"alla kostellattsj"one s"ulla s"ua d"Estra
```

In this work, the CLIPS corpus is used mainly for the evaluation of the FA task, but also as part of the evaluation of the ASR task.

Many publications reference the CLIPS corpus but few of them use it for FA[17, 22, 26, 106, 130].

5.1.3 Italian Parkinson's Voice and Speech (IPVS)

The Italian Parkinson's Voice and Speech (IPVS) dataset is an open access dataset, available under the Creative Commons Attribution License (CC BY 4.0)[49]. It was first presented in [50], which provides some details on how the dataset has been developed.

The dataset has been developed following a protocol illustrated in [50] which includes for each speaker the recording of

- 2 readings of phonemically balanced text
- repeated execution first of the syllable "pa" and then the syllable "ta"
- 2 phonation of each of the vowels ("a", "e", "i", "o" and "u")
- reading of phonemically balanced words
- reading of phonemically balanced phrases

In table 5.4 an overview of the phonemically balanced words, phrases and text.

The text in 5.4 is made to be long enough to require the patient to breathe with effort, to contain similar and complex phonetics close to each other so that they are harder to pronounce, and to require expression changes passing from one sentence to the next[50]. Phrases and words are also thought to be used to assess the degree of neurological control since they are constructed to stress all muscles involved in voice and speech production, requiring rapid and forceful movements[50].

The recordings in the dataset were performed with professional microphones in a quiet echo-free room, with a total of 65 Italian native speakers, divided into 3 categories[5, 50]:

- Young Healthy Control, 15 in number, age 20.8 ± 2.65
- Elderly Healthy Control, 22 of them, age 67.09 ± 5.16
- Patients with PD, age 67.21 ± 8.73, all 28 of them received their usual treatment for PD; the HY stage (section 2.5.2) for all the patients is less than 4, except for one at stage 5 and two at stage 4

It must be pointed out that most of the speakers are from the Bari area, in Apulia, a southern region of Italy[50].

In this work, the IPVS dataset is used for the evaluation of the ASR part of this work on the speech of patients with PD.

Text	*Il ramarro della zia.* Il papà (o il babbo come dice il piccolo Dado) era sul letto. Sotto di lui, accanto al lago, sedeva Gigi, detto Ciccio, cocco della mamma e della nonna. Vicino ad un sasso c'era una rosa rosso vivo e lo sciocco, vedendola, la volle per la zia. La zia Lulù cercava zanzare per il suo ramarro, ma dato che era giugno (o luglio non so bene) non ne trovava. Trovò invece una rana che saltando dalla strada finì nel lago con un grande spruzzo. Sai che fifa, la zia! Lo schizzo bagnò il suo completo rosa che divenne giallo come un taxi. Passava di lì un signore cosmopolita di nome Sardanapalo Nabucodonosor che si innamorò della zia e la portò con sé in Afghanistan
Phrases	• Oggi è una bella giornata per sciare. • Voglio una maglia di lana color ocra. • Il motociclista attraversò una strada stretta di montagna. • Patrizia ha pranzato a casa di Fabio. • Questo è il tuo cappello? • Dopo vieni a casa? • La televisione funziona? • Non posso aiutarti? • Marco non è partito. • Il medico non è impegnato.
Words	pipa, buco, topo, dado, casa, gatto, filo, vaso, muro, neve, luna, rete, zero, scia, ciao, giro, sole, uomo, iuta, gnomo, glielo, pozzo, brodo, plagio, treno, classe, grigio, flotta, creta, drago, frate, spesa, stufa, scala, slitta, splende, strada, scrive, spruzzo, sgrido, sfregio, sdraio, sbrigo, prova, calendario, autobiografia, monotono, pericoloso, montagnoso, prestigioso

Table 5.4. Phonemically balanced words, phrases and text in Italian used by [49, 50], the sentence in *italic* has also been used by [63]

5.2 Evaluation and metrics

The two main tasks characterizing this work are both evaluated at word and phoneme (or phone) level, for a total of four metrics:

- for the ASR task
  - Phoneme Error Rate (PER) (5.2.1)
  - Word Error Rate (WER) (5.2.2)
- while for the FA task
  - Phone Boundary Error (PBE) (5.2.3)

– Word Boundary Error (WBE) (5.2.4)

Furthermore, the Connectionist Temporal Classification (CTC) algorithm (5.2.5) is used as loss function during fine-tuning (5.4.1) and as a step to interpret the output of the model during inference.

5.2.1 Phoneme Error Rate (PER)

The Phoneme Error Rate (PER) is often used to evaluate the performance of ASR systems. In practice, it is a way to compare two strings of phonemic symbols and quantify how much they differ from one another.

Its definition can be derived from a normalized generalized Levenshtein distance[103, 188] with which it shares core concepts.

Using a notation similar to [188], given a symbols alphabet Σ, Σ^* being the set of strings over it and $\lambda \notin \Sigma$ being the null string; x_i denotes the ith element in the string $X = x_1x_2...x_n$ with $X \in \Sigma^*$.

An *elementary edit operation* can be defined as a pair $(a, b) \neq (\lambda, \lambda)$ with $a, b \in \Sigma \cup \{\lambda\}$, usually represented as $a \to b$, it can be of one of three types[188]

- insertion $\lambda \to a$, with $a \in \Sigma$
- substitution $a \to b$, with $a, b \in \Sigma$
- deletion $a \to \lambda$, with $a \in \Sigma$

Given two strings $X, Y \in \Sigma^*$, a sequence of elementary edit operations T_i to transform X into Y is denoted as edit transformation $T_{X,Y} = T_1T_2...T_l$[188]. Using a weight function $\gamma(a \to b)$, the weight of $T_{X,Y}$ is computed as[188]

$$\gamma(T_{X,Y}) = \sum_{i=1}^{l} \gamma(T_i)$$

a greater weight corresponds to a bigger penalty/cost or distance for an operation.

The Generalized Levenshtein Distance (GLD) among X and Y can then be defined as

$$GLD(X, Y) = \min\{\gamma(T_{X,Y})\} \tag{5.1}$$

and represents the difference between two strings, measured as the minimal number of insertions, deletions and substitutions needed to transform X into Y.

For GLD to be a distance function on Σ^* the weight function must satisfy $\forall a, b \in \Sigma \cup \{\lambda\}$ the following conditions[112, 188]

$$\gamma(a \to b) = \gamma(b \to a)$$

$$\gamma(a \to a) = 0 \tag{5.2}$$

$$k > 0, \gamma(a \to b) = k \Rightarrow a \neq b \tag{5.3}$$

5.1 can be normalized to obtain a normalized levenshtein distance or Normalized Edit Distance (NED)[112, 188]

$$NED(X,Y) = \min\left\{\frac{\gamma(T_{X,Y})}{L(T_{X,Y})}\right\} \tag{5.4}$$

where $L(T_{X,Y})$ is the number of elementary operations described by $T_{X,Y}$[112, 188]. Note that $L(T_{X,Y})$ also counts operations of the type $a \to a$ that according to 5.2 do not contribute to $\gamma(T_{X,Y})$ since $\gamma(a \to a) = 0$.

From 5.4, the most common formula for PER can be obtained setting $k = 1$ in 5.3, having Σ be the alphabet of phonemic symbols of interest and taking into account that an elementary edit operation of substitution $(a \to b)$ where $a = b$ is a "match" or "correct" symbol. Having X taking the role of the *hypothesis* string and Y the one of *reference* string, the symbols C, D, I and S respectively can be used to represent the number of correct symbols and the minimum number of deletions, insertions and substitutions that can be performed on the reference Y to obtain the hypothesis X. Using the symbols just defined it can then be seen that $\gamma(T_{X,Y}) = D + I + S$ and $L(T_{X,Y}) = C + D + I + S$ in 5.4. PER is usually normalized using the length N of the reference string Y, so in 5.4 $L(T_{X,Y}) = C + D + I + S$ is replaced with N to finally obtain

$$PER(X,Y) = PER(Hypothesis, Reference) = \frac{D+I+S}{N} = \frac{D+I+S}{C+D+S} \tag{5.5}$$

Considering a recording of the text "La macchina ha subito danni alla fiancata destra" from Common Voice, its phonological transcription computed as in 6.1 used as reference is R "la makina a subito danni alla fjankata dɛstra" and a prediction from the model or hypothesis could be H "alla makina subito daɲɲa alla fjankata destra". In this case it is possible to identify 2 insertions, 1 deletion, 4 substitutions and 33 hits. Using 5.5 and the data from this example, for this hypothesis-prediction association is possible to obtain

$$PER(H,R) = \frac{D+I+S}{C+D+S} = \frac{1+2+4}{33+1+4} = \frac{7}{38} \simeq 0.18.$$

5.2.2 Word Error Rate (WER)

The WER is a common metric for the evaluation of ASR or machine translation systems. It can be calculated in the same way as PER in 5.5

$$WER(Hypothesis, Reference) = \frac{D+I+S}{N} = \frac{D+I+S}{C+D+S} \tag{5.6}$$

where $Hypothesis, Reference \in \Sigma^*$, this time with Σ being the set of all words in a language.

In the context of this work, this metric has been used on words written with phonemes instead of graphemes.

Considering the same recording from the example in 5.2.1, with reference R "la makina a subito danni alla fjankata dɛstra" and prediction from the model H "alla makina

subito daɲɲa alla ʃjankata destra". In the case of WER, it is possible to identify 0 insertions, 1 deletion, 3 substitutions and 4 hits. Using 5.6 and the data from this example, for this hypothesis-prediction association is possible to obtain

$$WER(H,R) = \frac{D+I+S}{C+D+S} = \frac{1+0+3}{4+1+3} = \frac{4}{8} = 0.5.$$

5.2.3 Phone Boundary Error (PBE)

The Phone Boundary Error (PBE) is computed as the absolute value of the difference between a phone boundary from "manual"/reference alignment and the boundary proposed for the same phone by the FA system

$$PBE(b_h, b_r) = |b_h - b_r| \tag{5.7}$$

where b_h is a boundary from the hypothesis and b_r is the corresponding boundary in the reference. The value of the metric is expressed in a unit of measure of time, usually seconds or milliseconds.

This metric is usually calculated for all the boundaries in a reference sentence or the entire reference corpus. Given the two sequences of boundaries B_h and B_r, respectively hypothesis and reference, for the same sequence of phoneme labels, mean PBE can be computed as

$$PBE(B_h, B_r) = \frac{\sum_{i=1}^{|B_r|} |b_{hi} - b_{ri}|}{|B_r|} \tag{5.8}$$

In cases in which the sequence of phonemes in the hypothesis does not match exactly the sequence of the reference, it would not be correct to directly compare boundaries between the two. It is necessary to first "align the alignments" accounting for their position in the recording but also for the phone label matching, considering also possible insertions and deletions[117].

The alignment between the two sequences of boundaries is performed on a phoneme base and not for single boundaries; the left boundary, right boundary and phoneme label for each phoneme in the sequence are used for the alignment[117]. An overlap scoring function uses these parameters to classify a phoneme and its boundaries as a "match" or not[117].

Depending on the set of phonemes or phones on which the system or model to be evaluated has been trained and depending on the phonemes or phones used in the reference corpus, it is advisable to introduce a comparison or mapping function considering as matching/overlapping some sets of similar phones and allophones[117].

Moreover, it must be pointed out that the PBE metric is in some cases computed on a selection of boundaries not including all the ones present in the reference: [114] for examples considers only boundaries denotable as $.CVC$, $C.VC$, $CV.C$ and $CVC.$ (where C and V are respectively consonant and vowel, . represent a boundary).

5.2.4 Word Boundary Error (WBE)

The Word Boundary Error (WBE) is computed as the absolute value of the difference between a word boundary from "manual"/reference alignment and the boundary proposed for the same word by the FA system.

5.2.5 Connectionist Temporal Classification (CTC)

The CTC algorithm was introduced by [75] with the aim of enabling the training of RNNs for sequence-to-sequence tasks using real world input data with unsegmented labels[75]. The use of CTC enables thus the use of powerful sequence learners such as RNNs, and now the more recent Transformer based networks, on tasks like ASR or handwriting recognition that would have otherwise required pre-segmented training data, scarcely available and expensive to acquire[14, 75, 81].

The idea at the core of CTC is the interpretation of network outputs as a probability distribution over all possible label sequencesa, conditioned on a given input sequence[75]. From this distribution a differentiable objective function can be derived, that maximises the probabilities of correct labellings and allows training with backpropagation through time[75]. It was presented to solve the task of labelling unsegmented data sequences, also called *temporal classification* from which the name *Connectionist Temporal Classification*[75]. The temporal classification task was evaluated on a NED based metric[75] similar to PER (5.2.1).

In the context of this work, CTC allows the training of ASR systems using just audio and its unaligned transcript[81]. A more detailed description of the inner working of CTC follows, using the ASR task as an example.

Considering an audio input sequence $X = [x_1, x_2, \ldots, x_T]$ of length T, where each x_t is an audio frame, X is preprocessed or passed directly to the model M[75, 81]. Defining the set of symbols B as the output sequence alphabet, the set of symbols produceed as output of the softmax layer at the end of the CTC model is defined as $B' = B \cup \{\epsilon\}$, where the ϵ symbol is added as a special symbol representing empty output[13, 75, 181]. Each activation in the softmax layer represents the probability of observing the corresponding label or "blank" at a particular time[75]. Denoting the output of the model M for X as $A = M(X) = [a_1, a_2, \ldots, a_T]$, where $T = |X| = |A|$ is always the number of frames in X, each a_t can be seen as $a_t = [a_t^1, a_t^2, \ldots, a_t^k, \ldots, a_t^{|B'|}]$, having an element for each label k interpreted as the probability of observing it at frame t[75].

Using all the a_t^k terms is possible to obtain a distribution over B'^T, the set of sequences of lenght T over the alphabet B' defined as follow[75]

$$p(\pi|X) = \prod_{t=1}^{T} a_{\pi_t}^t, \forall \pi \in B'^T \tag{5.9}$$

where π, element of B'^T, is also called a *path* and is a possible alignment on X[13, 75].

On the a posteriori probabilities $p(\pi|X)$ just described, CTC applies a many-to-one mapping $g : B'^T \mapsto B^{\leq T}$, where $B^{\leq T}$ is the set of sequences over the alphabet B (with no ϵ) of length less than T[75]. The mapping g takes a path/alignment and maps it to a unaligned sequence of symbols, in some cases simple text, first removing consecutive

repeating symbols and then removing all the ϵ symbols from the alignment, obtaining a shorter sequence (length $\leq T$) and containing only symbols from B[13, 75]. For example, an alignment $hh\epsilon\epsilon ll\epsilon llo$ is transformed into $he\epsilon l\epsilon lo$ and then mapped to a text sequence $hello$[81]. Notice the role of the ϵ in preventing the merge of repeated symbols when its actually desirable to have multiple identical consecutive symbols, without the ϵ symbols the output text would have been $helo$[81]. The behavior of g corresponds to outputting a new symbol each time the model switches from predicting a label to no label and viceversa, or from a label to another[75]. In figure 5.1 a graphical representation of this example.

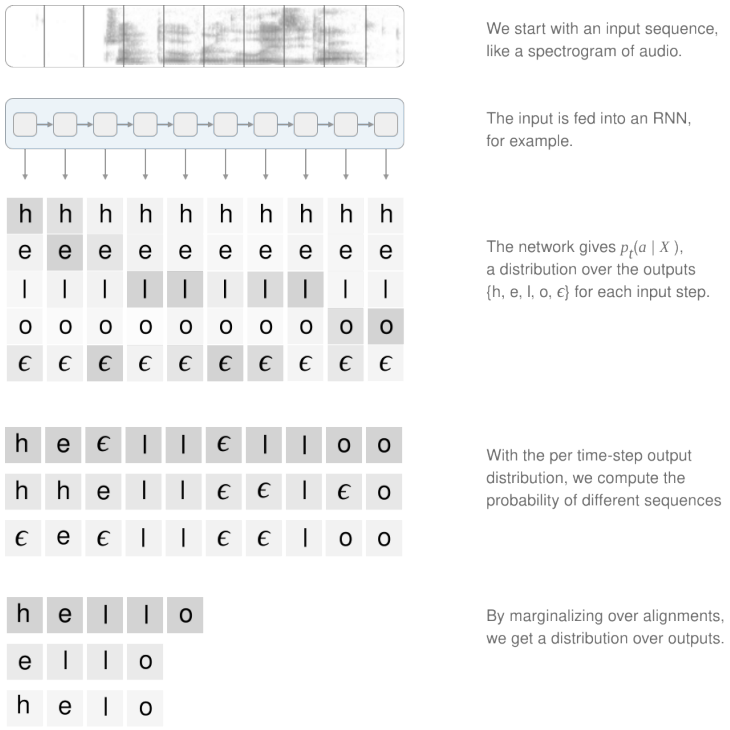

Figure 5.1. CTC example. Figure from [81]

The conditional probability of an element L of the set $B^{\leq T}$, obtained through g, can be computed summing the probabilities of all the alignments/paths mapping to it[75]

$$p(L|X) = \sum_{\pi \in g^{-1}(L)} p(\pi|X) \tag{5.10}$$

where the marginalization on the right hand side of the equation can be computed efficiently through dynamic programming[13, 81].

Given a training set D, during training with a CTC loss the parameters of the model are tuned to minimize[81]

$$\sum_{(X,Y) \in D} -\log p(Y|X) \tag{5.11}$$

where Y is the unaligned label for an audio X.

Inference, or decoding, is approximated with finding the most probable alignment and mapping it to a symbols/text sequence[13]

$$Y^* = g\left(\arg\max_A \prod_{t=1}^{T} p(a_t|X)\right) = g\left(\prod_{t=1}^{T} \arg\max_{a_t} p(a_t|X)\right). \tag{5.12}$$

5.3 Model

The model chosen to implement the solution to both the ASR and the FA tasks is XLS-R.

5.3.1 XLS-R

XLS-R is a "family" of large-scale cross-lingually pre-trained wav2vec 2.0 models[12] released in November 2021[12, 66]. XLS-R models can be used for speech processing tasks in many languages due to the cross-lignual representation learning, pre-trained models are available, and pre-training from scratch does not require labeled data. XLS-R models, like other wav2vec 2.0 models, according to [12, 14, 30] are composed of

- a **multi-layer convolutional feature encoder**, mapping raw audio to latent speech representations, it contains several blocks composed of temporal convolution followed by layer normalization and a GELU activation function; the feature encoder receptive field takes in input 400 samples at a time of raw audio, equivalent to 25ms with a sampling rate of 16KHz, and then proceeds with a stride of 20ms
- a **quantization module**, that discretizes the output of the feature encoder to a finite set of speech representations in self-supervised training[3]
- a **context network**, implemented using the Transformer architecture[171] from the NLP model BERT[48], it takes in input latent speech representations and outputting context representations; differently from BERT it uses a convolutional layer to perform relative positional embedding instead of absolute positional embedding

Figure 5.2 shows a block diagram of the model during pre-training.

Overall, the model starts from audio, produces continuous speech representations and from these derive context representations; then self-attention captures dependencies over the entire sequence of latent representations end-to-end[14].

Three variants have been published with 317 million, 965 million and 2162 million parameters[12]. Models so large became common in NLP with strong results on established benchmarks when trained on adequately large datasets, spanning billions of documents[12].

[3]The concept of *supervision* or *supervised learning*, in the context of machine learning, refers to a paradigm for the learning of mapping functions from labeled data, that is data for which, for each sample composed of a number of features or data points, there is an annotation or label associcated. *Unsupervised learning* is instead characterized by the use of only features or data points, and no labels. *Self-supervision or Self-supervised learning* is a paradigm characterized by the use of unlabeled data first on a pretext task to initialize network Weights followed by supervised or unsupervised learning on the actual task[189].

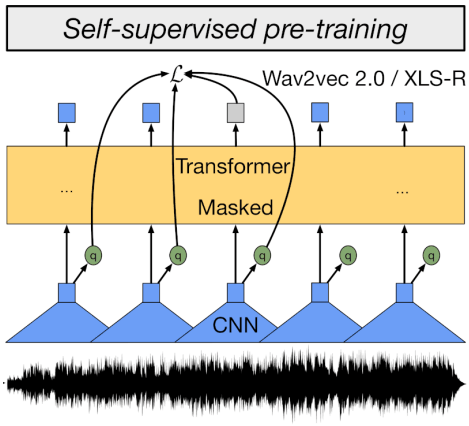

Figure 5.2. **Self-supervised pre-training model blocks diagram**. Figure adapted from [12]. In black unlabeled speech input, in blue the multi-layer convolutional feature encoder, in green the quantization module and in yellow the context network.

XLS-R uses a similar approach as NLP models, its name is inspired by the NLP model XLM-R[12], and applies it to speech. The models are pretrained on 436K hours of publicly available data, across 128 different languages, from the following corpora[12]:

- 372K hours of unlabeled data from VoxPopuli[176] in 23 European languages from parliamentary speech
- 50K hours of data from Multilingual Librispeech[144] in 8 European languages
- 7K hours from CommonVoice[8] v6.1, the December 2020 release, which covers 60 languages
- 6.6K hours of data in 107 languages based on YouTube content from VoxLingua107[169]
- about 1K hours of data from BABEL[68] of conversational telephone speech in 17 African and Asian languages

Not all the languages are represented in the same way in the obtained dataset: 24 of them have more than 1K hours each, 17 have between 1K and 100 hours and 88 have less than 100 hours each[12].

To learn from unlabeled data in the obtained dataset a self-supervised learning paradigm has been used[14]. This paradigm has been found to be particularly successful for NLP, it is usually followed by fine-tuning using a supervised or semi-supervised approach on labeled data[14]. Self-supervised learning is characterized by the definition and use of *pretext* tasks which do not need labeled data but to be solved require high level semantic understanding of the input data[189], and are thus useful to learn a general data representations from unlabeled data[14]. For example, in computer vision, a commonly used task is the prediction of correct image rotation, under the assumption that the model

will be required to learn and understand useful and general concepts about the objects depicted in the image (such as their location in the image, their type, and their pose) to be able to predict if the rotation is correct[71].

In the case of XLS-R, and wav2vec2.0 models, the self-supervised pre-training uses a task similar to the "masked LM" from BERT[12, 14, 48]. The model is required to solve a contrastive task, that is one with a loss function that encourages the hidden representations of the same object to be closer together and those of different objects to be further apart[164]. More specifically, at training time, some segments of the outputs of the feature encoder are masked, and the model is asked to identify the correct quantized latent audio representation in a set of distractors for each masked segment[12, 14, 30]. Moreover, pre-training is performed on multiple languages simultaneously, multilingual batches are formed using a distribution with a parameter to control the importance given to high-resource versus low-resource languages[12, 30].

The pre-training phase just described is followed by fine-tuning using labeled data. XLS-R if fine-tuned and evaluated on a set of different tasks, with appropriate specific datasets, to demonstrate the ability of generalization of the pre-trained model[12]. The tasks are: Automatic Speech Translation (AST), ASR, language identification and speaker identification. For the AST a decoder[4] Transformer network is stacked on top of the model, for the language identification and speaker identification task a linear layer acting as classifier is added on top of the pretrained model[12]. For the ASR task a linear classification output layer is added on top too, randomly initialized and with classes for each character/grapheme or phoneme in the dataset used, training is then done using CTC loss[75]; a language model is added on top of the classifier for some of the datasets used for evaluation but not for CommonVoice[12, 14]. Weights of the feature encoder are not updated at fine-tuning time, at first only the added classifier layer is trained, after a while also the Transformer network is updated[12, 14, 30].

Figure 5.3 shows a block diagram of the model during fine-tuning.

The benefits of the multilingual pre-training presented by [12, 30] led to XLS-R being the state of the art on CommonVoice and other ASR datasets, at the time it was published[12].

The wav2vec2.0 architecture has been chosen as base for implementation of work presented in this text, more specifically its XLS-R pre-trained version, due to its strong performance on CommonVoice, in several languages including Italian, upon fine-tuning[12].

5.4 Methodologies and architecture

This work experiments with the use of an XLS-R/wav2vec2.0 based E2E model to perform ASR and FA on both normophonic and non-normophonic speech.

A pre-trained model is first fine-tuned to perform ASR outputting Italian phonemes, and it is evaluated in terms of WER and PER on normophonic and non-normophonic speech.

[4]Referring to the encoder-decoder architecture, see [171] for more details on a Transformer based encoder-decoder architecture

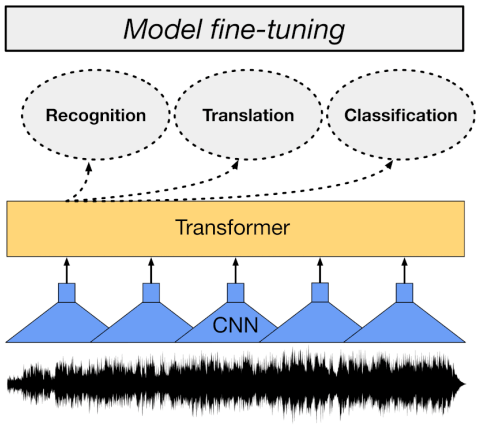

Figure 5.3. **Fine-tuning model blocks diagram**. Figure adapted from [12]. In black unlabeled speech input, in blue the multi-layer convolutional feature encoder, in yellow the context network, while the gray circles represent the blocks added for the different tasks on which the model can be fine-tuned.

The output probabilities of the fine-tuned model are then leveraged to perform FA, the results are evaluated on normophonic speech in terms of WBE and PBE.

5.4.1 ASR

For the ASR task the selected pre-trained XLS-R model is fine-tuned using the same architecture used by [12] and a similar methodology. It is worth to point out that [12] uses CommonVoice as a "few-shots"[5] learning benchmark, thus performing training and evaluation for each language only on a very small subset of the available material from CommonVoice[12]. In this work instead, the entirity of the training split for Italian CommonVoice is used for training and test split for evaluation, after the prepocessing described in part II. During the fine-tuning the training loss, validation loss and validation WER are computed and monitored. The CTC loss is used as it is done by [12] for ASR.

The fine-tuned model is later evaluated on CommonVoice, CLIPS and IPVS computing WER and PER.

5.4.2 Forced Alignment (FA)

To extract the alignment information from the output probabilities of the XLS-R based model from 5.4.1, fine-tuned on ASR with a CTC loss, we use the method proposed by [98] with an implementation similar to [84], discussed in further details in part II.

[5]Few-shots learning is a machine learning method aiming at training models to predict the correct class of instances when a small number of examples are available in the training dataset[183].

The transcribed phonemes and the frame-wise posteriors probabilities, obtained as output of the model, are used conjunctly to produce the maximum joint probability of alighment of the text until a certain character up to a certain audio frame[98]. The maximum joint probability at a point is computed comparing the two possible transitions, the one to the blank symbol or the one to the next character[98]. Backtracking on these maximum joint probabilities, from the most probable temporal position of the last phoneme in the transcription, the phoneme-wise alignment is obtained[98].

The resuts of this method are evaluated on CLIPS in terms of PBE and WBE, with a strategy inspired by [117].

See figure 5.4 for an overview of the steps in the architecture adopted and the data flow which characterizes the experiments described in part II.

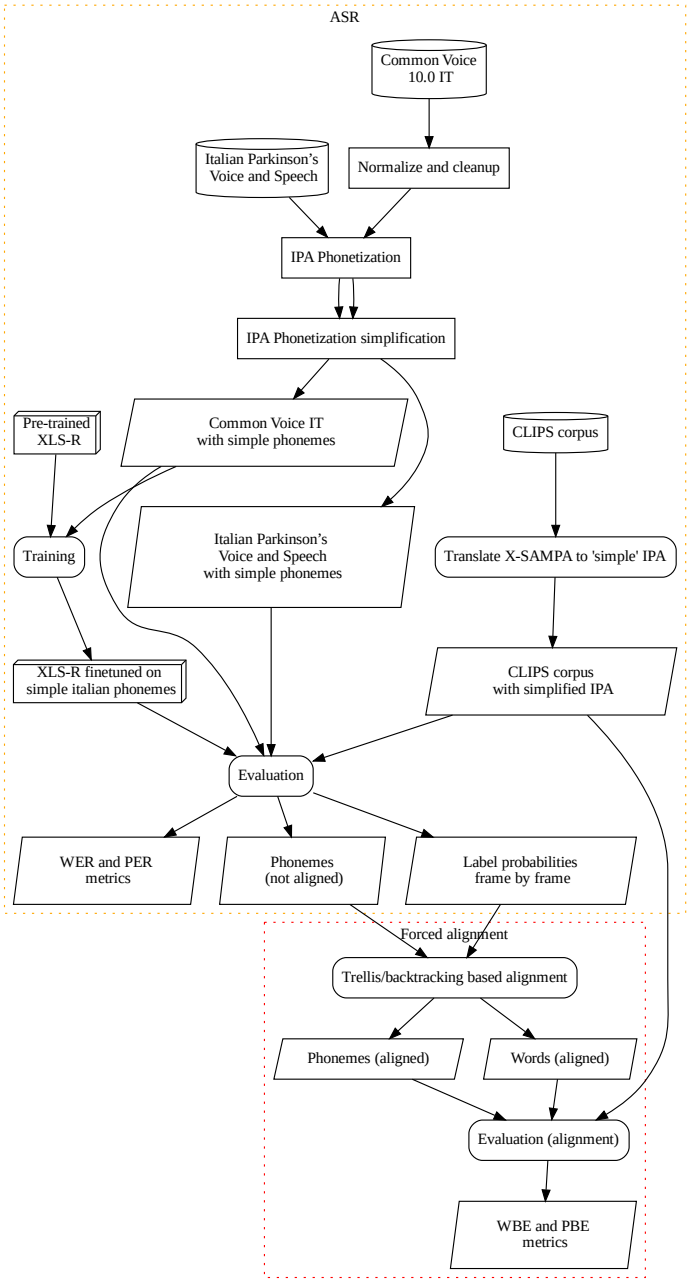

Figure 5.4. Overall architecture; cylinder-shaped nodes represent corpora or datasets, parallelogram-shaped nodes represent data in general, box-shaped nodes represent models, rectangular-shaped nodes represent a specific step or process, more specifically rectangular-shaped nodes with rounded corners are complex processes with several steps

Part II

Experiments and solution implementation

Chapter 6

Data preparation and preprocessing

During data preparation and preprocessing particular attention has been dedicated to making each step and the overall process as much reproducible as possible, with this intent, each operation on data has been done through python scripts that can function as precise descriptions of the procedure.

6.1 Common Voice

The Italian sub-corpus of Common Voice 10.0 is downloaded from [122], to start the download is necessary to insert a valid email address. The download consists of one compressed archive in ".tar.gz" format of about 8.07 GigaBytes. Extracting the archive we are presented with a folder "clips", containing all the recordings in MP3 format, and a group of "TSV" files. Each line in these "TSV" files references a recording in the "clips" folder and provides the label and metadata for it.

A python script has been developed to take as input one of these "TSV" files, read all of its lines and produce in output a "CSV" file, which contains rows only for the lines that passed all the filtering, normalization and preprocessing operations applied by the script.

The lines from the TSV file are divided into chunks of configurable size, each chunk is then processed by a worker from a pool of configurable size. Each worker executes on each row in the chunk the following actions in order

- remove the "accents" field
- normalize label graphemes and filtering
- graphemes to IPA phonemes
- normalize and simplify IPA phonemes

If an error is detected in one of these actions the row is discarded. Out of the 149591 labeled clips in the training split and 14974 in the test split respectively 144541 and 14439 pass all the steps, while 5050 and 535 are discarded.

6.1.1 Normalize label graphemes and filtering

For this operation, the script mostly leverages a validator for Italian from the python package commonvoice-utils[1], many other languages are supported too.

The validator allows only a list of graphemes to be present in the label text, while removing or replacing others for normalization.

If the label text contains a grapheme for which no rule is specified, the word containing it is not an Italian word and the corresponding recording is discarded.

6.1.2 Graphemes to IPA phonemes

The Common Voice corpus does not provide phonetic or phonemic labeling for its material. With the purpose of obtaining phonemic labels from the label text two software were individuated and evaluated: espeak-ng and phonetisaurus. The results of the evaluation led to the use of espeak-ng.

espeak-ng or eSpeak NG is a compact[2] open source software TTS synthesizer for Linux, Windows, Android and other operating systems. It supports more than 100 languages and accents, it is available as a command line program and, among other features, it can be used to translate text into IPA phonemes[58]. It was chosen over phonetisaurus because of its ability to identify words that do not belong to the Italian language even though all its graphemes are usable in Italian.

In this operation, espeak-ng is used through the python package py-espeak-ng[3], more specifically its function for grapheme to phoneme translation. The output of this function is then checked for words marked as belonging to another language, if such markings are found the sentence is discarded. If the sentence is not discarded the output goes through a mapping function that replaces some phonemes with their most common and similar allophone, see table 6.1.

espeak phoneme	allophone
ʊ	u
ɪ	i
ɾ	r
ɹ	r

Table 6.1. espeak-ng phonemes conversion to allophone

[1] https://github.com/ftyers/commonvoice-utils

[2] The program and its data, including many languages, totals about few Mbytes[58]

[3] https://pypi.org/project/py-espeak-ng

6.1.3 Normalize and simplify IPA phonemes

Only a subset of IPA phones and phonemes has been chosen to be used as the dictionary for input and output to and from the model. The selected subset of phonemes is the one that is most often recognized as the phonemic inventory of Standard Italian and is illustrated in table 6.2

Allowed IPA symbols
a, e, i, o, u, ɛ, ɔ, p, b, t, d, k, g, f, v, s, z, m, n, l, r, z, j, w, ʃ, ɲ, ʎ, ŋ, ŋ

Table 6.2. Allowed IPA phonemes, 29 symbols

Note how phones that in Italian are usually considered a single phoneme such as t͡s, d͡z, t͡ʃ and d͡ʒ here are decomposed into two removing the linking, adopting a simplification approach similar to the one proposed in [116]. Following [116], diacritics and prosody symbols are removed, see table 6.3. Sentences that, after the removal of diacritics and prosody symbols still contain symbols that are not present in table 6.2 are discarded.

Diacritics	ː long, i.e. /ɑː/ ˑ half long, i.e. /ɑˑ/ "̆" extra short, i.e. /i̯/ "̯" non syllabic, i.e. /i̯/ "͡" linking, i.e. /d͡ʒ/ "X" linking, i.e. /dXʒ/ "‿" linking, i.e. /d‿ʒ/ "̩" syllabic, i.e. /n̩/ "̪" dental
Prosody	ˈ ˌ .

Table 6.3. Removed IPA symbols for diacritics and prosody

6.2 Italian Parkinson's Voice and Speech

The Italian Parkinson's Voice and Speech is downloaded from [49] as a ".zip" archive. The archive is extracted. Based on information in [49] and [50], a script has been realized to output a ".csv" file that contains, for each audio file, its path, category(young healthy control, elderly healthy control, and patient with PD), person name, recording type (see 5.1.3), prompt text, phonemes from prompt text.

The prompt text is obtained from a code in each file name which according to documents in [49] corresponds to a prompt from [50]. There are prompt texts and not transcriptions, so the content of the recordings may differ from it. As for Common Voice no phonemic transcription is provided for the prompts, the same process described for

Common Voice in 6.1.1, 6.1.2 and 6.1.3 is applied to obtain the phonemic transcription of each prompt text.

6.3 Corpora e Lessici di Italiano Parlato e Scritto (CLIPS) corpus

The CLIPS corpus is available in the "Private Area" of its website. Following the registration process, it becomes possible to start browsing, through the website, the folders and files composing the dataset. There is no direct way to download the entire dataset as a single archive.

6.3.1 Download and cleaning

As part of this work, a simple crawler has been developed in python using Selenium and Firefox webdriver to browse the corpus section of the website in-depth, collecting the URLs for all the files belonging to the dataset. In this way, a list of 129878 unique URLs, one URL per file, has been obtained and saved to a file. The size of the list is relatively close to 128457, the number provided in attachment 10 of [159].

A note on the CLIPS website reports that "some folders" contain zip files corresponding to the entire contents of the same name directory on the same path. However, further investigation into this note has led to finding that

- this ".zip" archives are provided only for some labeled parts of the corpus
- the files contained in the ".zip" archives are not the most updated version of the file provided by the website, the same file out of the archive seems to be newer or to contain corrections that are not present in the archive

For this reason, it was not possible to download only the ".zip" archives from the obtained list, instead, all the files from the list had to be downloaded.

Another python script has been developed to take as input the file containing the list and download the files from all the URLs of the list. The script is written to accept the "JSESSIONID" cookie for an authenticated user on the corpus website and use it to execute multiple downloads in parallel using a pool of 8 downloaders, the downloaded files are organized by the script with the same folder structure used by the corpus website.

Having downloaded all the corpus files, aggregated durations for the material have been computed using another custom-made python script. A level of the directory structure of the corpus divides recordings with segment labels, "etichettato" folders, from the other material, "corpus" folders. However, recordings in the "etichettato" folders are duplicated in "corpus" folders, the script takes this into account computing durations separately for both. The results from this script slightly differ from the ones advertised in [159] and are illustrated in 6.4 and 6.5.

As described in 5.1.2 for each audio file several files are present for labels at different levels. To simplify the handling of the dataset a script has been developed to write a JSON file containing a list of items, one per audio file, each containing the path to the file and all the labels for it. The script scans in depth the path in which the corpus

Dialogico	RTV	Letto	Ortofonico	Telefonico
48:17:25.060	16:51:37.331	16:04:21.752	3:42:29.317	16:43:32.772

Table 6.4. CLIPS corpus overall recordings total duration computed per sub-corpus from downloaded files, expressed as hours:minutes:seconds.milliseconds

Dialogico	RTV	Letto	Ortofonico	Telefonico
6:21:16.887	1:23:18.721	1:19:03.810	0:33:18.979	0:35:18.444

Table 6.5. CLIPS corpus labeled recordings total duration computed per sub-corpus from downloaded files, expressed as hours:minutes:seconds.milliseconds

has been saved, looking for ".wav" files in the "etichettato" directories. For each ".wav" file found the script looks for ".phn", ".st_", ".wr_", ".wrd", and ".std" files to extract the respective labels.

This script has also been used to check for inconsistencies in the corpus. Of the inconsistencies spotted in this way some were corrected with a human-supervised fuzzy-finder and some were corrected manually, a "patch file" has been generated both to have a precise description of the changes that have been made and to make it possible to apply them automatically on a version of the corpus downloaded from the corpus website. The categories of inconsistencies that have been found in this way are:

- inconsistent file names: when files for labels have names that slightly differ from the corresponding ".wav" files; these inconsistencies have been fixed by uniforming the label files name to the ".wav" file name
- missing files: when some label files are missing for a ".wav" file; some of these inconsistencies have been fixed by reconstructing the content of the missing label file from the other label files and the audio

Not all the inconsistencies could easily be fixed, in those cases the associated audio file and labels were discarded. Table 6.6 shows the changes in total recordings durations after discarding said inconsistencies.

While collecting and organizing the labels for each audio file the script checks for the presence of the needed files with appropriate names and, having found the needed file for a given label type, checks that the content of the label file complies with specifications from [158]. A parser and a validator have been written and used for every label of interest, some inconsistencies have been found also in this way and the corresponding material has been discarded.

Particular effort has been put in the conversion from SAMPA and X-SAMPA encoded labels to a subset of IPA, further details about it are described in sub-section 6.3.2.

6.3.2 X-SAMPA phones to IPA phonemes

The script writing the JSON file for the dataset leverages a custom parser to translate X-SAMPA phones directly into simplified IPA as defined in 6.1.3.

Dialogico	RTV	Letto	Ortofonico	Telefonico
6:20:39.557	1:18:38.431	1:18:55.268	0:33:18.979	0:35:00.708

Table 6.6. CLIPS corpus labeled recordings total duration computed per sub-corpus from downloaded files, expressed as hours:minutes:seconds.milliseconds

The parser has been written using ANother Tool for Language Recognition (ANTLR)[4]. ANTLR is a parser generator for reading, processing, executing, or translating structured text or binary files, widely used to build languages, tools, and frameworks[135].

From a language grammar, a formal language description, it generates a parser for that language that can automatically build parse trees, data structures representing how a grammar matches a given input[135]. It also automatically generates tree walkers that can be used to visit the nodes of those trees to execute application-specific code[135].

Following the specifics in [158], [182] and [185] a grammar and a lexicon have been written in a ".g4" ANTLR 4 grammar file.

The lexicon maps every X-SAMPA symbol to a descriptive name, these descriptive names are used by grammar rules to describe how the corresponding symbols can be combined to describe a phone. The resulting grammar counts more than 16 rules.

ANTLR 4 is used to generate, from this ".g4" file, python3 code for a lexer, a parser and a basic parse tree visitor. The basic autogenerated parse tree visitor class is extended in another python3 class to implement the custom logic necessary for the translation to simplified IPA.

[4] https://www.antlr.org/

Chapter 7

Experiments

This chapter illustrates the activities of fine-tuning that have been performed and the setup for their evaluation. The chapter is organized in two sections, 7.1 and 7.2, respectively describing the fine-tuning of the XLS-R model and the implementation of the FA algorithm. Each of the two section contains a subsection detailing the evaluation process, 7.1.2 for the ASR part and 7.2.2 for the FA one.

7.1 End-to-end ASR with XLS-R

These experiments have been run on a Google Colab Pro virtual machine, with 54 Gigabytes of RAM and a dual-core CPU, taking advantage of Nvidia GPUs such as the Tesla V100 with 16 Gigabytes of video RAM, to accelerate training and evaluation time.

7.1.1 Fine-tuning

The goal of this activity is to obtain a model that takes in input an audio segment containing speech and outputs its transcription in a reduced set of selected IPA phonemes. See figure 7.1 for an overview of the operations composing this experiment.

To save training time without compromising on performance, the XLS-R model is not trained from scratch: a pre-trained model is chosen and then fine-tuned. Three different versions of the model, with different numbers of parameters, have been released by [12] and are available for download[1]

- `Wav2Vec2-XLS-R-300M`, with 300 million parameters
- `Wav2Vec2-XLS-R-1B`, with 1 billion parameters
- `Wav2Vec2-XLS-R-2B`, with 2 billion parameters

[1] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

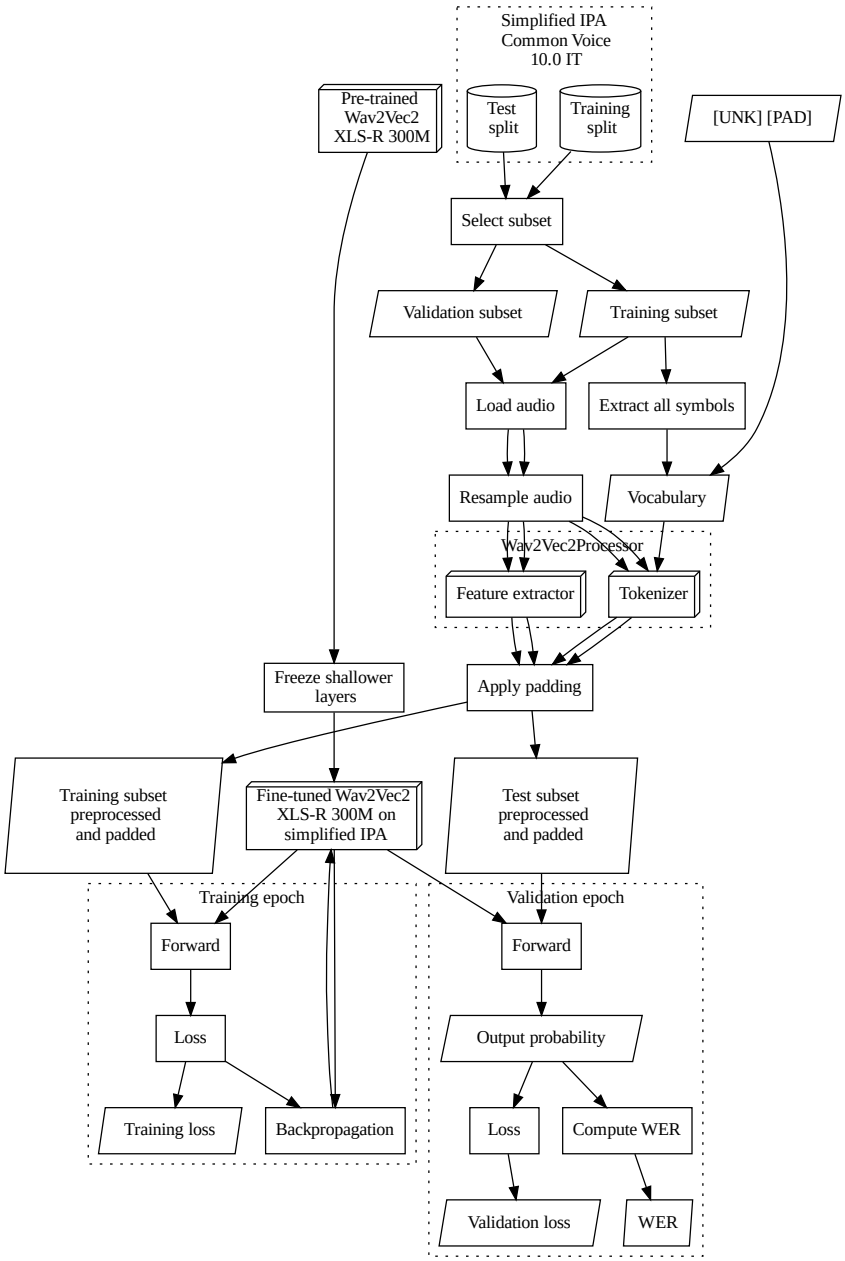

Figure 7.1. Flow of data in preprocessing, training and validation during the fine-tuning of a pre-trained XLS-R model; cylinder-shaped nodes represent corpora or datasets, parallelogram-shaped nodes represent data in general, box-shaped nodes represent models, rectangular-shaped nodes represent a specific step or process

The model with 300 million parameters, `Wav2Vec2-XLS-R-300M`, has been chosen for the experiment.

The modified version of the Common Voice corpus, prepared as described in 6.1, is used for fine-tuning the model. A subset of 50000 and 5000 items respectively from "training" and "test" splits is selected and carried through preprocessing.

The preprocessing consists of loading the MP3 audio files given the path for each selected item from the JSON file of the corpus, and then resample it to 16KHz so that the input to the model has a uniform sample rate. Then, of all the attributes available for each item, only the resampled audio and "simplified" IPA label are carried on for the fine-tuning of the model.

A Wav2Vec2 feature extractor is configured and used, this performs pre-processing audio files to Log-Mel Spectrogram features, normalization and padding. It is configured to work with a 16KHz sampling rate, a feature dimension of 1, padding with zeros and returning the attention mask.

A dictionary of symbols is obtained from the labels of the selected items of the dataset, it is checked that the dictionary contains all the allowed symbols defined in 6.1.3. Note that in the subset selected for training the ŋ symbol is not represented. However, it is still added to the dictionary, but it won't be used. Symbols for "padding" and "unknown" are added to the dictionary, and the result is saved in a file and is then used to configure the tokenizer from `Wav2Vec2CTCTokenizer`.

The training and test subsets are prepared to be used by the `Trainer`: for each item in a subset the audio is passed through the feature extractor and the simplified IPA label through the tokenizer. The output of these two steps are kept for training and validation, while the other attributes are discarded.

The model thus takes as input batches of N items containing 3 attributes each `input_values`, `input_length`, and `label`.

The pretrained model is adapted to the custom vocabulary size and its shallower layers, responsible for feature extraction, are frozen, since the goal is to only learn to map the current internal representation of phones to the selected subset of "simple" IPA symbols.

The `Trainer` class from HuggingFace transformers is configured to perform training and evaluation loops. The experiment carries out 10 epochs, with batch size of 16, learning rate of 3e-4, 500 warmup steps and floating point precision of 16 bits.

The model uses the CTC loss, discussed in 5.2.5. Together with the loss, the Trainer is configured to compute the WER metric for each evaluation step.

7.1.2 Evaluation

The model obtained in 7.1.1 is evaluated against the entire test split of the Italian subset of Common Voice, the entirety of IPVS, and a large portion of CLIPS, all in the version that has been previously prepared to include simplified IPA labels (as explained in 6). Differently from the evaluation epochs performed during fine-tuning, the evaluation epoch is now performed on the entire corpus chosen, there is no subset selection.

The finetuned model, feature extractor and tokenizer from 7.1.1 are loaded.

With a procedure similar to 7.1.1, the three corpora are preprocessed and prepared for the model: resampling, feature extraction, tokenization, normalization and padding

are performed.

An HuggingFace `Trainer` instance is configured to perform an evaluation "epoch" for each of the three prepared corpora.

The `Trainer` instance is configured to compute WER and PER metrics, results are presented and discussed in 8.1.

7.2 FA with XLS-R

The FA task experiment uses the CTC-segmentation algorithm from [98] with an implementation similar to [84]. Some key differences with [84] are introduced by the use of a different model, the use of phonemes as output instead of graphemes, the different metrics computed, the integration with the HuggingFace library and several changes needed to perform FA on an entire corpus and correctly aggregate values for the metrics.

The experiments in this section were run on a laptop with 32 Gigabytes of RAM and a 14-cores Intel CPU, taking advantage of a Nvidia GeForce RTX 3060 GPU with 6 Gigabytes of GDDR6 video RAM, to accelerate inference and evaluation time.

The first step for FA is to load the finetuned model, feature extractor and tokenizer from 7.1.1.

Then follow a series of step similar to the ones described in 7.1.1: each item in the corpus goes through resampling, feature extraction, normalization, and padding. In this case it is not needed to tokenize the "transcription" label, since it is not used in this kind of evaluation.

7.2.1 FA of a clip

Given a clip from the corpus, its preprocessed audio is passed as input to the fine-tuned model which produces frame-wise output probabilities in the form of a matrix with time frames on one axis and phoneme labels on the other axis. Figure 7.2 show the frame-wise output probabilities for an excerpt of a recording from the IPVS corpus (the original audio is longer and would have resulted in less readable plots).

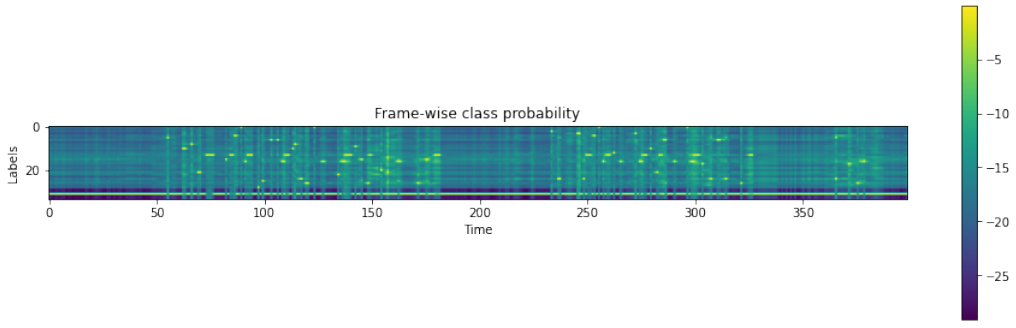

Figure 7.2. Frame-wise probabilities obtained as output of the model, given as input a portion of a recording from IPVS

The `argmax` function is applied frame by frame and its result is passed to the tokenizer decode function to obtain a prediction of the transcription of the recording. In the case of the example in figure 7.2: "ɔdʒi e una bɛlla dʒornata per ʃiare voʌo una maʌa di lana kolor okra".

In this case however, the output probabilities will be used for more than predicting the transcription. Using the frame-wise output probabilities and the transcription of the recording obtained from the model it is possible to generate the alignment probability. A 2D matrix is defined, with time on an axis and labels from the transcription on the other, also called *trellis*. Using t to denote the index on the time axis, j to denote the index on the labels axis, and k to denote the trellis matrix, the elements of k can be computed as[84]

$$k_{(t+1,j+1)} = \max(k_{(t,j)}p(t+1, c_{j+1}), k_{(t,j+1)}p(t+1, repeat)) \tag{7.1}$$

where c_j is the label at index j, $p(t, c_j)$ represents the probability of label c_j at time step t and $repeat$ is the "blank" token from CTC (5.2.5)[84].

The main characteristic of the trellis matrix, observable from equation 7.1, is that $k_{(t+1,j+1)}$ is the maximum between two values, corresponding to the two only possible choices during alignment[84]

- staying at the same label, represented by $k_{(t,j+1)}p(t+1, repeat)$,
- and transitioning to the next label, represented by $k_{(t,j)}p(t+1, cj+1)$.

Moreover, the use of $\max$ implies that the more probable step choice for $k_{(t+1,j+1)}$ is taken[84].

Figure 7.3 represent the trellis computed for the example transcription and the frame-wise output probabilities from figure 7.2.

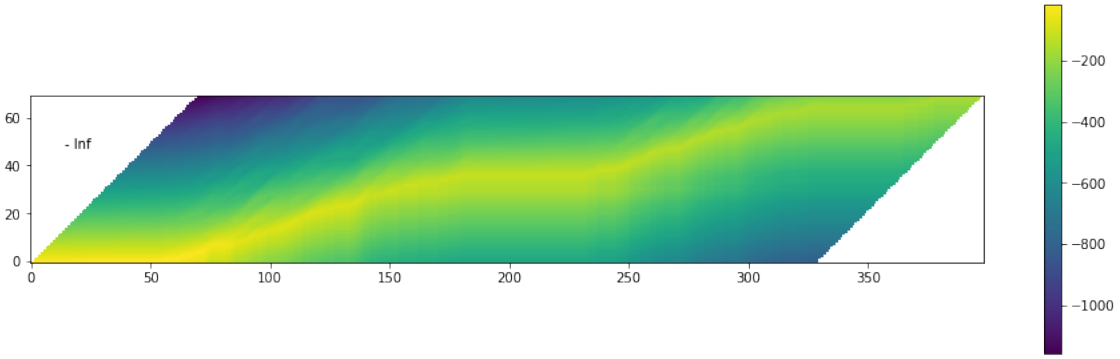

Figure 7.3. Trellis obtained from frame-wise probabilities and transcription predicted by the ASR CTC-based model

Given the trellis matrix, the most likely path on it, corresponding to the alignment, is found through backtracking. That is, starting from the last label index at its time step of highest probability, the trellis is traversed going back in time.

Traversing, the algorithm chooses if[84]

- stay at index j with label c_j, based on the post-transition probability $k_{t-1,j}p(t-1, c_{j-1})$
- or transition to the following label index $j-1$, based on the post-transition probability $k_{t-1,j-1}p(t-1, repeat)$

The trellis matrix is used for path finding but the confidence for each segment in the alignment is computed based on the frame-wise probability[84].

Figure 7.4 show the most likely path for the trellis matrix in figure 7.3 and the frame-wise probabilities in figure 7.2.

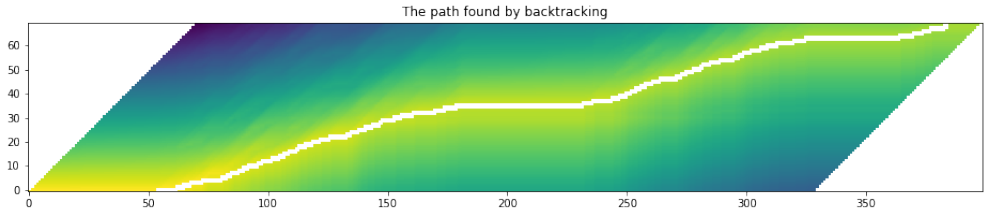

Figure 7.4. Most probable path on the trellis obtained from frame-wise probabilities and transcription predicted by the ASR CTC-based model

The points from the path constitute the alignment, but at this stage it still contains repetitions. The points corresponding to repetitions are merged into segments, averaging the probability from each point being merged[84]. With this last step the alignment is finally computed.

Figure 7.5 shows the alignment overlaid on the trellis path, together with segment and point probabilities.

Figure 7.6 overlays the alignment on a graph of the audio track.

7.2.2 Evaluation

A graphical overview of the evaluation steps for the FA task evaluation is provided in figure 7.7

The alignment strategy and the model are evaluated, for the FA task, on the CLIPS corpus. The sequence of segments (produced as described in 7.2.1) describing the model hypothesis for start, end, and label for each phoneme in a recording must be aligned to the reference sequence to compute the absolute distances between corresponding boundaries. To obtain this alignment an approach similar to [117] has been adopted.

After loading and preprocessing the CLIPS corpus (see section 6.3), a python script has been written in order to iterate the procedure over each clip and:

- perform the steps described in 7.2.1 to obtain the alignment segments;
- accumulate values to compute PBE, having aligned hypothesis and reference similarly to [117];
- merge alignment segments into words;

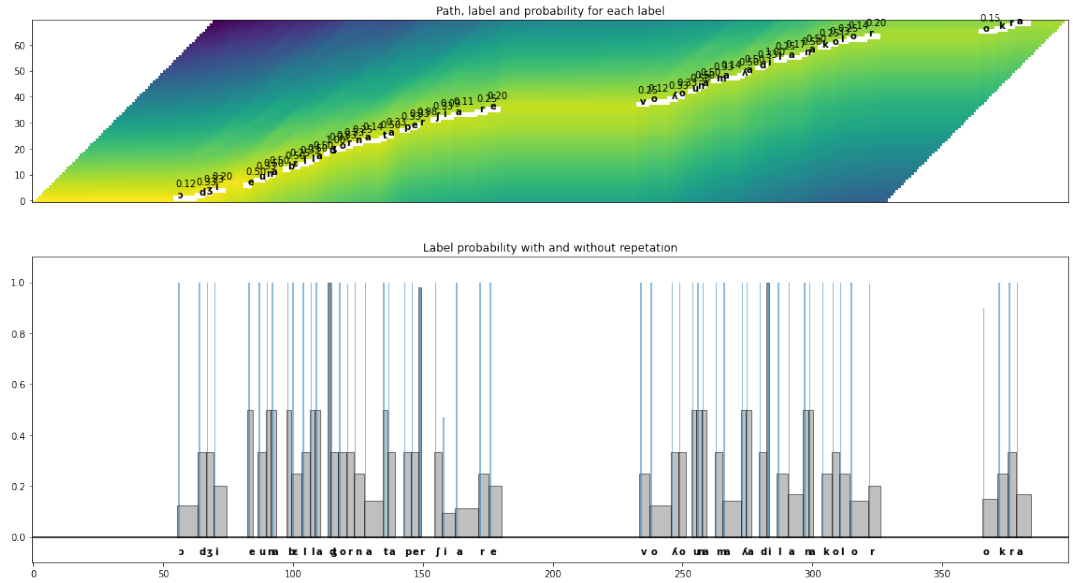

Figure 7.5. Most probable path on the trellis with labels and corresponding probabilities

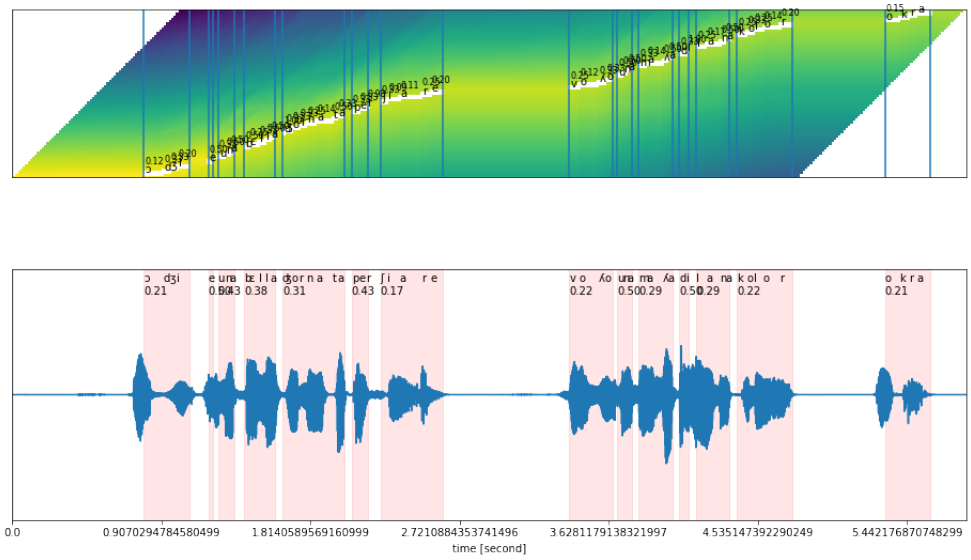

Figure 7.6. Labels overlaid on corresponding audio track

- accumulate values to compute WBE, having aligned hypothesis and reference similarly to [117].

Finally, once every clip in the corpus has been aligned and the necessary data has been collected, PBE and WBE have been computed.

Results for the evaluation are presented and discussed in 8.

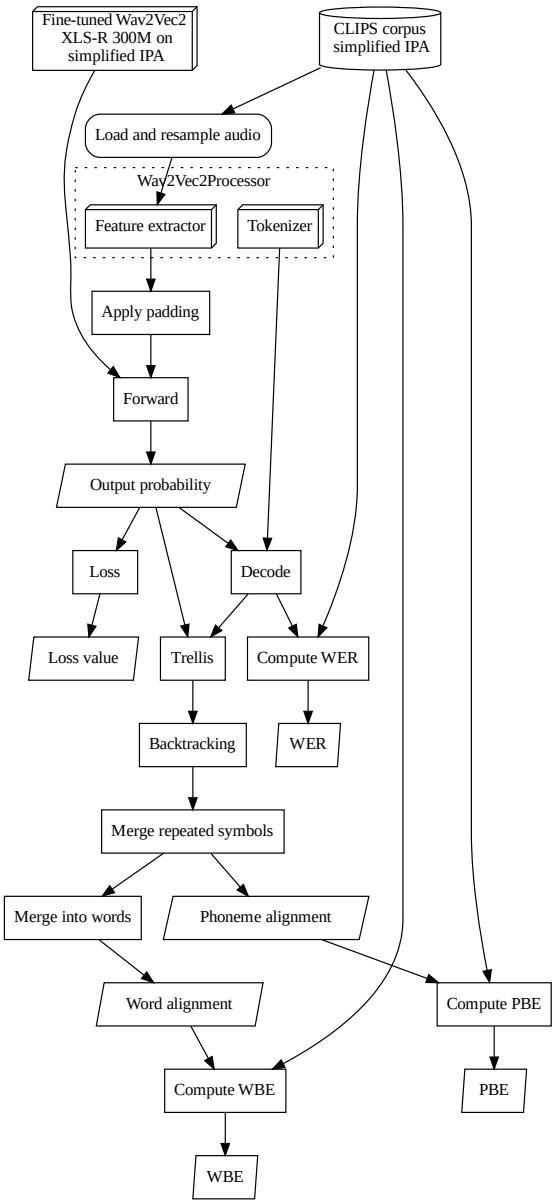

Figure 7.7. Flow of data in preprocessing and validation during FA with fine-tuned XL-S-R model; cylinder-shaped nodes represent corpora or datasets, parallelogram-shaped nodes represent data in general, box-shaped nodes represent models, rectangular-shaped nodes represent a specific step or process

Chapter 8

Results and discussion

8.1 End-to-end ASR

The results for the ASR task are presented in table 8.1

Corpus	PER [%]	WER [%]
Common Voice	3.28	13.31
CLIPS	10.39	32.62
IPVS	12.33	34.59

Table 8.1. Overview results ASR

The results on Common Voice indicate that our fine-tuned model outperforms the models from Babu et al. [12], the same size model for them had a 4.9 PER while their larger models obtained a 3.5 PER[12]. This is due to the fact that Babu et al. in [12] performed fine-tuning on Italian Common Voice simulating a few-shot scenario, using only one hour of training data.

Usually, the performance of an ASR model on Common Voice is evaluated using CER, but since our model outputs phonemes and the hypothesis for a clip is evaluated on the result of applying G2P conversion to the transcription from Common Voice, comparing the results we obtained for PER with state-of-the-art CER is imprecise. However, considering the CER results on Italian for XLS-R and UniSpeech-ML from [53](equal to 3.9% and 2.4%, respectively) we can see that our result are very close to the state-of-the-art.

Concerning the WER on Common Voice, a consideration similar to the previous about CER/PER applies: models from other works produce transcription in graphemes and thus WER is computed on words written with graphemes, while in our case the model produces phonemic transcriptions and the "words" that are used to compute WER are written with phonemes. XLS-R is not evaluated in terms of WER on Common Voice by [12], but they compute WER for orthographic words on the Italian subset of Multilingual LibriSpeech (MLS)[144]. The best XLS-R model obtained a 12.1% WER in a few shot scenario, while the state-of-the-art performance using the entire MLS Italian subset is

found to have 10.5% WER. Also in this case the results from our model are close to the state-of-the-art.

The evaluation on IPVS is the first one on an out-of-domain corpus, given that the model is fine-tuned only on Common Voice.

As can be seen in table 8.1, despite an acceptable performance reduction due to the fact that the model is not personalized and has not been trained on pathological voices, the model proved efficient also in presence of dysartric speech (as in PD patients).

The CLIPS corpus is also out-of-domain, but the evaluation on it adds more complications:

- The phonemic transcription for the recordings in Common Voice and IPVS are obtained through the same G2P system (as described in 6.1.2), while CLIPS already included manually annotated phonemic transcriptions that have been used after some preprocessing (described in 6.3.2). The G2P system can adopt conventions different from the ones adopted by the human operators, being the model trained on the transcriptions from the G2P system its has the same conventions and biases.
- Differently from Common Voice and IPVS it contains also spontaneous speech

Due to these complications, the results on CLIPS are inevitably worse than on Common Voice.

8.2 Forced Alignment (FA)

For the evaluation of the FA task, 491394 phoneme boundaries and 119148 word boundaries from the alignments predicted by the model have been matched successfully with boundaries in the reference alignments from CLIPS.

For each couple of predicted-reference boundary the absolute boundary error has been computed. From this, PBE (5.2.3) and WBE (5.2.4) have been evaluated, obtaining 49ms and 63ms, respectively.

The values distribution for absolute boundary error is further analyzed in table 8.2, where the percentage of boundaries meeting different tolerances in presented, and in figures 8.1 and 8.2, showing the distribution of the absolute boundary errors for phoneme and word boundaries, respectively.

	Absolute boundary error [ms]							
Level	**≤10**	**≤20**	**≤30**	**≤40**	**≤50**	**≤75**	**≤100**	**≤500**
Phoneme	21%	39%	54%	66%	75%	87%	92%	99%
Word	19%	35%	47%	58%	67%	82%	88%	99%

Table 8.2. Distribution of absolute boundaries errors

The results of the FA evaluation are in line with the results obtained by systems with similar architectures on other corpora, such as [98] and [107]. Despite a small difference can be appreciated between the current results and the state of the art (i.e. MFA[114] and NeuFA[104]), it should be noted that the CLIPS corpus contains more complex speech

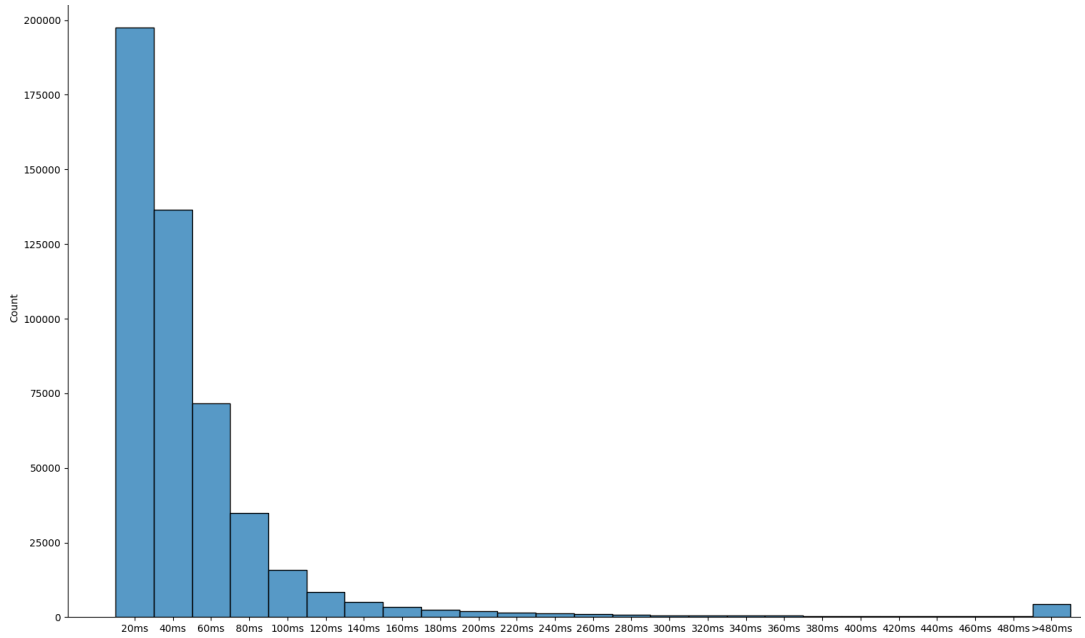

Figure 8.1. Distribution of absolute phone boundaries error

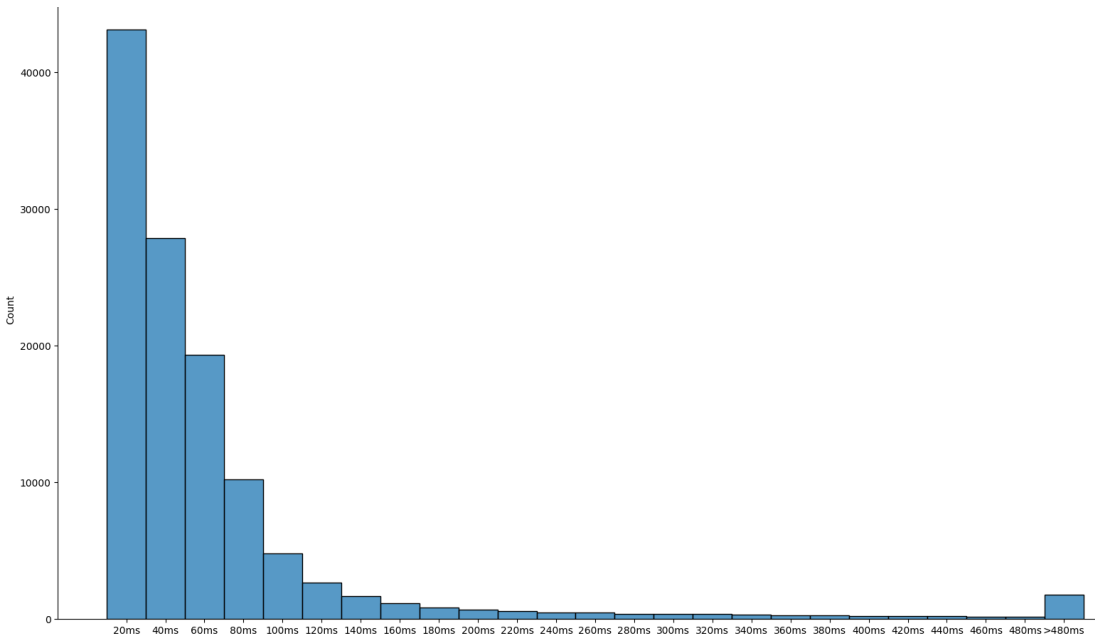

Figure 8.2. Distribution of absolute word boundaries error

examples compared to Buckeye[138], and the system we developed does not need a transcription to be provided.

The latter one is particularly important for the use case envisioned in this work, since recordings often contain speech not present in the transcription. Conventional FA systems instead can easily be confused by the presence of utterances absent in the provided transcription.

Unfortunately, it was not possible to evaluate the performance of the FA system on speech from PD patients because IPVS does not offer any transcription or segmentation, only the prompts used for each recording.

Part III

Future work and conclusions

Chapter 9

Contributions, future work and conclusions

This last chapter lists the contributions made by this work (9.1), analyzes some possible future developments (9.2) and summarizes the conclusions drawn from the experiments and their results (9.3).

9.1 Contributions

The contributions of the work presented in this master's degree thesis are mainly related to the use of state-of-the-art methodologies for ASR and FA on the Italian language and speech from PD patients, taking into account the requirements for their use in an AVSCA system for the diagnosis and monitoring of PD. More specifically, these include:

1. An integrated system for ASR and/or FA that can be used as a building block for AVSCA systems
2. An analysis of the state-of-the-art for FA on Italian
3. An analysis of the state-of-the-art for ASR on speech from speaker with speech disorders
4. First use of E2E models for FA on Italian
5. Code to download and parse the CLIPS corpus for its use with various Python frameworks for deep learning

9.2 Future work

Given the time constraints of a master's degree thesis work, some refinements to the proposed methodology could not be realized but are identified by the author as necessary next step to improve the system:

- improvements to the fine-tuning methodology
  - include the use of data augmentation methods, such as SpecAugment[133]
  - include other methods to reduce the "catastrophic forgetting" phenomena[1] while fine-tuning to enable the possibility of "unfreezing" other layers
- bridge the domain gap between Common Voice and CLIPS for more accurate evaluation of the FA task
  - improvements to the G2P system and to the script used to convert CLIPS labels from X-SAMPA to IPA
  - fine-tuning on both Common Voice and a portion of CLIPS
- leverage model confidence to determine which segments to provide to the downstream task (evaluation or AVSCA system)
- evaluate the FA capabilities of the system on speech from PD patients

After these refinements, it would be possible to explore the expansion paths that were considered while defining the architecture:

- support for multiple languages
  - XLS-R, and other E2E models for ASR, have been pre-trained on multiple languages
  - the transcription representation in IPA phonemes can represent multiple languages
  - fine-tuning for ASR could be performed simultaneously for more than one language
  - a multilingual system can be used to leverage more corpora of PD speech from different languages
- semi-supervised or self-supervised learning on speech data from PD patients with no segmentation or transcription

Moreover, during the development of this thesis work, as observed in the literature review, other E2E models like UniSpeech-ML[177] and Whisper[149] have been found to have better performance on Italian[53] and better performance with distorted audio. The use of these newer models is possible with few changes to the current architecture and could lead to better performances.

Other interesting developments are possible with more relevant changes to the current work and include:

- Methodology change: Even if its output can be used to perform FA, the model on which the current system is based is designed for ASR; it could be replaced with an E2E model like NeuFA[104], designed for FA, a better support for the boundaries positioning can be added on top of an existing E2E ASR model architecture.

[1] When an ANN abruptly forgets previously learned information upon learning new information[115]

- Architecture Change: The current architecture leverages a speaker-independent model, but given the results obtained by personalized ASR models on disordered speech (4.2.5) and the limited amount of data needed for the personalization, it would be interesting to adopt a similar approach for a FA system.

9.3 Conclusions

This master's degree thesis explored state-of-the-art methods for ASR and FA, described an architecture combining them in a system which can be used as a block in an AVSCA system, detailed an implementation of said combined ASR-FA system, and evaluated its performance on three different corpora.

The implemented system allows for phoneme level segmentation, robust to repetitions, skipping of words or syllables, mispronunciation, and insertion of superfluous phonemes which occur frequently in the speech of PD patients. The robustness is obtained also by removing the need for a predefined prompt or known transcription, thanks to the ASR capabilities of the model. The model used in the system is pre-trained on unlabeled speech data, available in relatively larger quantities than labeled data. With this in mind, the objectives defined in the introduction (1.1) can be considered achieved.

The results confirm that E2E models offer state-of-the-art performance for ASR.

Concerning the FA task, the use of an E2E model improves robustness to unexpected utterances, but currently causes a small reduction in the accuracy of boundaries positioning.

Bibliography

[1] *aeneas: automagically synchronize audio and text*. URL: https://www.readbeyond.it/aeneas/ (visited on Nov. 29, 2022).

[2] Hanan Aldarmaki et al. "Unsupervised Automatic Speech Recognition: A review." In: *Speech Communication* 139 (Apr. 2022), pp. 76–91. ISSN: 01676393. DOI: 10.1016/j.specom.2022.02.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167639322000292 (visited on Aug. 23, 2022).

[3] *alfabeto fonetico in "Enciclopedia dell'Italiano"*. URL: https://www.treccani.it/enciclopedia/alfabeto-fonetico_%28Enciclopedia-dell%27Italiano%29/ (visited on Apr. 3, 2022).

[4] Federica Amato et al. "An algorithm for Parkinson's disease speech classification based on isolated words analysis." In: *Health Information Science and Systems* 9.1 (Dec. 2021), p. 32. ISSN: 2047-2501. DOI: 10.1007/s13755-021-00162-8. URL: https://link.springer.com/10.1007/s13755-021-00162-8 (visited on May 28, 2022).

[5] Federica Amato et al. "Speech Impairment in Parkinson's Disease: Acoustic Analysis of Unvoiced Consonants in Italian Native Speakers." In: *IEEE Access* 9 (2021), pp. 166370–166381. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3135626. URL: https://ieeexplore.ieee.org/document/9650872/ (visited on May 13, 2022).

[6] B. Angelini et al. "Automatic segmentation and labeling of English and Italian speech databases." In: *3rd European Conference on Speech Communication and Technology (Eurospeech 1993)*. 3rd European Conference on Speech Communication and Technology (Eurospeech 1993). ISCA, Sept. 22, 1993, pp. 653–656. DOI: 10.21437/Eurospeech.1993-158. URL: https://www.isca-speech.org/archive/eurospeech_1993/angelini93_eurospeech.html (visited on Nov. 28, 2022).

[7] *APASCI – ELRA Catalogue*. URL: https://catalogue.elra.info/en-us/repository/browse/ELRA-S0039/ (visited on Nov. 28, 2022).

[8] Rosana Ardila et al. *Common Voice: A Massively-Multilingual Speech Corpus*. Mar. 5, 2020. arXiv: 1912.06670[cs]. URL: http://arxiv.org/abs/1912.06670 (visited on Aug. 31, 2022).

[9] Arnold E. Aronson and Diane Bless. *Clinical Voice Disorders*. Google-Books-ID: kmugBjlqGBkC. Thieme, Jan. 1, 2011. 346 pp. ISBN: 978-1-58890-661-8.

[10] Patricia Ashby. *Understanding: Phonetics*. Understanding Language series. 244 pp. ISBN: 978-0-340-92827-1.

[11] Frederico A.C. Azevedo et al. "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain." In: *Journal of Comparative Neurology* 513.5 (2009). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.21 pp. 532–541. ISSN: 1096-9861. DOI: 10.1002/cne.21974. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.21974 (visited on June 8, 2022).

[12] Arun Babu et al. "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale." In: *arXiv:2111.09296 [cs, eess]* (Dec. 16, 2021). arXiv: 2111.09296. URL: http://arxiv.org/abs/2111.09296 (visited on Apr. 10, 2022).

[13] Alexei Baevski et al. *Unsupervised Speech Recognition*. May 2, 2022. arXiv: 2105.11084[cs,eess]. URL: http://arxiv.org/abs/2105.11084 (visited on Oct. 6, 2022).

[14] Alexei Baevski et al. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." In: *arXiv:2006.11477 [cs, eess]* (Oct. 22, 2020). arXiv: 2006.11477. URL: http://arxiv.org/abs/2006.11477 (visited on May 7, 2022).

[15] Ladan Baghai-Ravary and Steve W. Beet. *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*. SpringerBriefs in Electrical and Computer Engineering. New York, NY: Springer New York, 2013. ISBN: 978-1-4614-4573-9 978-1-4614-4574-6. DOI: 10.1007/978-1-4614-4574-6. URL: http://link.springer.com/10.1007/978-1-4614-4574-6 (visited on Nov. 5, 2022).

[16] Fabio Ballati, Fulvio Corno, and Luigi De Russis. "Assessing Virtual Assistant Capabilities with Italian Dysarthric Speech." In: *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '18: The 20th International ACM SIGACCESS Conference on Computers and Accessibility. Galway Ireland: ACM, Oct. 8, 2018, pp. 93–101. ISBN: 978-1-4503-5650-3. DOI: 10.1145/3234695.3236354. URL: https://dl.acm.org/doi/10.1145/3234695.3236354 (visited on Nov. 30, 2022).

[17] Pierpaolo Basile et al., eds. *EVALITA. Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 7 December 2016, Naples*. Accademia University Press, 2016. ISBN: 978-88-99982-55-3. DOI: 10.4000/books.aaccademia.1899. URL: http://books.openedition.org/aaccademia/1899 (visited on May 7, 2022).

[18] Gaetano Berruto and Massimo Cerruti. *La linguistica. Un corso introduttivo*. 2° edizione. Torino: UTET Università, Apr. 3, 2017. 334 pp. ISBN: 978-88-6008-483-5.

[19] Pier Marco Bertinetto and Michele Loporcaro. "The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome." In: *Journal of the International Phonetic Association* 35.2 (Dec. 2005), pp. 131–151. ISSN: 0025-1003, 1475-3502. DOI: 10.1017/S0025100305002148. URL: https://www.cambridge.org/core/product/identifier/S0025100305002148/type/journal_article (visited on July 9, 2022).

[20] Brigitte Bigi. "A Phonetization Approach for the Forced-Alignment Task in SPPAS." In: *Human Language Technology. Challenges for Computer Science and Linguistics*. Ed. by Zygmunt Vetulani, Hans Uszkoreit, and Marek Kubis. Vol. 9561. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 397–410. ISBN: 978-3-319-43807-8 978-3-319-43808-5. DOI: 10.1007/978-3-319-43808-5_30. URL: http://link.springer.com/10.1007/978-3-319-43808-5_30 (visited on Nov. 24, 2022).

[21] Brigitte Bigi. *SPPAS the automatic annotation and analysis of speechand analysis of speech*. 2011. URL: http://www.sppas.org/doc/SPPAS-Documentation.pdf (visited on Nov. 27, 2022).

[22] Brigitte Bigi. "The SPPAS participation to Evalita 2011." In: (), p. 21.

[23] Per Borghammer. "How does parkinson's disease begin? Perspectives on neuroanatomical pathways, prions, and histology: Where Does Parkinson's Disease Begin?" In: *Movement Disorders* 33.1 (Jan. 2018), pp. 48–57. ISSN: 08853185. DOI: 10.1002/mds.27138. URL: https://onlinelibrary.wiley.com/doi/10.1002/mds.27138 (visited on June 8, 2022).

[24] Heiko Braak et al. "Staging of brain pathology related to sporadic Parkinson's disease." In: *Neurobiology of Aging* 24.2 (Mar. 2003), pp. 197–211. ISSN: 01974580. DOI: 10.1016/S0197-4580(02)00065-9. URL: https://linkinghub.elsevier.com/retrieve/pii/S0197458002000659 (visited on June 7, 2022).

[25] Zeinab Breijyeh and Rafik Karaman. "Comprehensive Review on Alzheimer's Disease: Causes and Treatment." In: *Molecules* 25.24 (Dec. 8, 2020), p. 5789. ISSN: 1420-3049. DOI: 10.3390/molecules25245789. URL: https://www.mdpi.com/1420-3049/25/24/5789 (visited on June 3, 2022).

[26] Francesco Cangemi et al. "AUTOMATIC SPEECH SEGMENTATION FOR ITALIAN (ASSI): TOOLS, MODELS, EVALUATION AND APPLICATIONS." In: AISV Italian national conference. Lecce, Italy, Jan. 26, 2011, p. 8.

[27] Honglei Chen and Beate Ritz. "The Search for Environmental Causes of Parkinson's Disease: Moving Forward." In: *Journal of Parkinson's Disease* 8 (s1 Dec. 18, 2018). Ed. by Patrik Brundin, J. William Langston, and Bastiaan R. Bloem, S9–S17. ISSN: 18777171, 1877718X. DOI: 10.3233/JPD-181493. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JPD-181493 (visited on June 5, 2022).

[28] Frank C. Church. "Treatment Options for Motor and Non-Motor Symptoms of Parkinson's Disease." In: *Biomolecules* 11.4 (Apr. 20, 2021), p. 612. ISSN: 2218-273X. DOI: 10.3390/biom11040612. URL: https://www.mdpi.com/2218-273X/11/4/612 (visited on June 6, 2022).

[29] Jennifer Cole. "Prosody in context: a review." In: *Language, Cognition and Neuroscience* 30.1 (Feb. 7, 2015), pp. 1–31. ISSN: 2327-3798, 2327-3801. DOI: 10.1080/23273798.2014.963130. URL: http://www.tandfonline.com/doi/abs/10.1080/23273798.2014.963130 (visited on June 20, 2022).

[30] Alexis Conneau et al. *Unsupervised Cross-lingual Representation Learning for Speech Recognition*. Dec. 15, 2020. arXiv: 2006.13979[cs,eess]. URL: http://arxiv.org/abs/2006.13979 (visited on Sept. 15, 2022).

[31] Marshall Cavendish Corporation. *Mammal Anatomy: An Illustrated Guide*. Google-Books-ID: mTPI_d9fyLAC. Marshall Cavendish, 2010. 292 pp. ISBN: 978-0-7614-7882-9.

[32] Piero Cosi, Daniele Falavigna, and Maurizio Omologo. "A preliminary statistical evaluation of manual and automatic segmentation discrepancies." In: *2nd European Conference on Speech Communication and Technology (Eurospeech 1991)*. 2nd European Conference on Speech Communication and Technology (Eurospeech 1991). ISCA, Sept. 24, 1991, pp. 693–696. DOI: 10.21437/Eurospeech.1991-183. URL: https://www.isca-speech.org/archive/eurospeech_1991/cosi91_eurospeech.html (visited on Nov. 27, 2022).

[33] Piero Cosi et al. "Building Resources for Verbal Interaction Production and Comprehension within the ALIZ-E Project." In: *Il farsi e il disfarsi del linguaggio. Acquisizione, mutamento e destrutturazione della struttura sonora del linguaggio* 1 (Collana Studi AISV. Curatori del volume 1: Mario Vayra, Cinzia Avesani e Fabio Tamburini Dec. 18, 2015), pp. 455–458. DOI: 10.17469/O2101AISV000029. URL: https://doi.org/10.17469/O2101AISV000029 (visited on Nov. 27, 2022).

[34] PIERO COSI et al. "CHILDIT2 – A New Children Read Speech Corpus." In: *LA FONETICA NELL'APPRENDIMENTO DELLE LINGUE* 2 (2016), pp. 269–273. DOI: 10.17469/O2102AISV000016. URL: https://doi.org/10.17469/O2102AISV000016 (visited on Nov. 27, 2022).

[35] Piero Cosi et al. "KALDI: Yet Another ASR Toolkit? Experiments on Italian children speech." In: *Il farsi e il disfarsi del linguaggio. Acquisizione, mutamento e destrutturazione della struttura sonora del linguaggio* 1 (Collana Studi AISV. Curatori del volume 1: Mario Vayra, Cinzia Avesani e Fabio Tamburini Dec. 18, 2015), pp. 429–438. DOI: 10.17469/O2101AISV000027. URL: https://doi.org/10.17469/O2101AISV000027 (visited on Nov. 27, 2022).

[36] Cosi, Piero, Cutugno, Francesco, and Galatà, Vincenzo. "Forced Alignment on Children Speech." In: *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014 9-11 December 2014, Pisa*. pisa university press, 2014. DOI: 10.12871/clicit2014223. URL: http://clic2014.fileli.unipi.it/proceedings/Proceedings-CLICit-2014.pdf (visited on Aug. 31, 2022).

[37] David Crystal and David Crystal. *A dictionary of linguistics and phonetics*. 6th ed. The language library. OCLC: ocn187300284. Malden, MA ; Oxford: Blackwell Pub, 2008. 529 pp. ISBN: 978-1-4051-5296-9 978-1-4051-5297-6.

[38] Francesco Cutugno, Antonio Origlia, and Dino Seppi. "EVALITA 2011: Forced Alignment Task." In: *Evaluation of Natural Language and Speech Tools for Italian*. Ed. by Bernardo Magnini et al. Red. by David Hutchison et al. Vol. 7689. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 305–311. ISBN: 978-3-642-35827-2 978-3-642-35828-9. DOI: 10.

1007/978-3-642-35828-9_33. URL: http://link.springer.com/10.1007/978-3-642-35828-9_33 (visited on May 7, 2022).

[39] Francesco Cutugno and Renata Savy. "CLIPS. Diatopic, diamesic and diaphasic variations in spoken Italian." In: *PROCEEDINGS FROM THE CORPUS LINGUISTICS CONFERENCE SERIES*. 5th Corpus Linguistic Conference. Vol. 2009. Liverpool, 2009, #213,1–24. URL: https://ucrel.lancs.ac.uk/publications/cl2009/213_FullPaper.doc.

[40] Francesco Cutugno, Dino Seppi, and Antonio Origlia. "EVALITA 2011 Forced Alignment on Spontaneous Speech." EVALITA 2011 Workshop. Rome, Jan. 24, 2012. URL: https://www.evalita.it/wp-content/uploads/2021/11/Origlia.pdf (visited on Nov. 28, 2022).

[41] *Cytoplasm*. Genome.gov. URL: https://www.genome.gov/genetics-glossary/Cytoplasm (visited on June 5, 2022).

[42] *Cátedra RTVE de la Universidad de Zaragoza*. URL: http://catedrartve.unizar.es/rtvedatabase.html (visited on Nov. 29, 2022).

[43] Khashayar Dashtipour et al. "Speech disorders in Parkinson's disease: pathophysiology, medical management and surgical approaches." In: *Neurodegenerative Disease Management* 8.5 (Oct. 2018), pp. 337–348. ISSN: 1758-2024, 1758-2032. DOI: 10.2217/nmt-2018-0021. URL: https://www.futuremedicine.com/doi/10.2217/nmt-2018-0021 (visited on June 21, 2022).

[44] Jacob Oliver Day and Stephen Mullin. "The Genetics of Parkinson's Disease and Implications for Clinical Practice." In: *Genes* 12.7 (June 30, 2021), p. 1006. ISSN: 2073-4425. DOI: 10.3390/genes12071006. URL: https://www.mdpi.com/2073-4425/12/7/1006 (visited on June 5, 2022).

[45] Luigi De Russis and Fulvio Corno. "On the impact of dysarthric speech on contemporary ASR cloud platforms." In: *Journal of Reliable Intelligent Environments* 5.3 (Sept. 2019), pp. 163–172. ISSN: 2199-4668, 2199-4676. DOI: 10.1007/s40860-019-00085-y. URL: http://link.springer.com/10.1007/s40860-019-00085-y (visited on Nov. 25, 2022).

[46] *Definitions of Communication Disorders and Variations*. American Speech-Language-Hearing Association. Publisher: American Speech-Language-Hearing Association. 1993. URL: https://www.asha.org/policy/rp1993-00208/ (visited on June 19, 2022).

[47] Kris Demuynck et al. "SPRAAK: An Open Source "SPeech Recognition and Automatic Annotation Kit"." In: (), p. 1.

[48] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. May 24, 2019. arXiv: 1810.04805[cs]. URL: http://arxiv.org/abs/1810.04805 (visited on Oct. 4, 2022).

[49] Giovanni Dimauro and Francesco Girardi. "Italian Parkinson's Voice and Speech." In: (June 11, 2019). Publisher: IEEE Type: dataset. DOI: https://dx.doi.org/10.21227/aw6b-tg17. URL: https://ieee-dataport.org/open-access/italian-parkinsons-voice-and-speech (visited on Sept. 10, 2022).

[50] Giovanni Dimauro et al. "Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System." In: *IEEE Access* 5 (2017), pp. 22199–22208. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2017.2762475. URL: http://ieeexplore.ieee.org/document/8070308/ (visited on May 13, 2022).

[51] E. R. Dorsey et al. "Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030." In: *Neurology* 68.5 (Jan. 30, 2007), pp. 384–386. ISSN: 0028-3878, 1526-632X. DOI: 10.1212/01.wnl.0000247740.47667.03. URL: https://www.neurology.org/lookup/doi/10.1212/01.wnl.0000247740.47667.03 (visited on June 4, 2022).

[52] E. Ray Dorsey et al. "Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016." In: *The Lancet Neurology* 17.11 (Nov. 2018), pp. 939–953. ISSN: 14744422. DOI: 10.1016/S1474-4422(18)30295-3. URL: https://linkinghub.elsevier.com/retrieve/pii/S1474442218302953 (visited on June 3, 2022).

[53] Jonatas Dos Santos Grosman. "ASSESSMENT OF FINE-TUNING ON END-TO-END SPEECH RECOGNITION MODELS." DOUTOR EM CIÊNCIAS - INFORMÁTICA. Rio de Janeiro, Brazil: PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO, Sept. 15, 2022. DOI: 10.17771/PUCRio.acad.61086. URL: http://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=61086@2 (visited on Nov. 26, 2022).

[54] Brittany N. Dugger and Dennis W. Dickson. "Pathology of Neurodegenerative Diseases." In: *Cold Spring Harbor Perspectives in Biology* 9.7 (July 2017), a028035. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a028035. URL: http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a028035 (visited on June 2, 2022).

[55] Sean R Eddy. "What is a hidden Markov model?" In: *Nature Biotechnology* 22.10 (Oct. 2004), pp. 1315–1316. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt1004-1315. URL: http://www.nature.com/articles/nbt1004-1315 (visited on Nov. 19, 2022).

[56] Elisabeth Engl and David Attwell. "Non-signalling energy use in the brain: Non-signalling energy use in the brain." In: *The Journal of Physiology* 593.16 (Aug. 15, 2015), pp. 3417–3429. ISSN: 00223751. DOI: 10.1113/jphysiol.2014.282517. URL: http://doi.wiley.com/10.1113/jphysiol.2014.282517 (visited on June 8, 2022).

[57] Michael G. Erkkinen, Mee-Ohk Kim, and Michael D. Geschwind. "Clinical Neurology and Epidemiology of the Major Neurodegenerative Diseases." In: *Cold Spring Harbor Perspectives in Biology* 10.4 (Apr. 2018), a033118. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a033118. URL: http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a033118 (visited on June 2, 2022).

[58] *espeak-ng/espeak-ng: eSpeak NG is an open source speech synthesizer that supports more than hundred languages and accents.* URL: https://github.com/espeak-ng/espeak-ng (visited on Sept. 2, 2022).

[59] *EVALITA 2014*. evalita. URL: https://www.evalita.it/campaigns/evalita-2014/ (visited on Nov. 27, 2022).

[60] *evalita – Evaluation of NLP and Speech Tools for Italian*. evalita. URL: https://www.evalita.it/ (visited on Nov. 27, 2022).

[61] Stanley Fahn, C. David Marsden, and Peter Jenner. *Recent Developments in Parkinson's Disease*. Ed. by Paul Teychenne. New York: Raven Pr, Jan. 1, 1986. 375 pp. ISBN: 978-0-88167-132-2.

[62] Gunnar Fant. *Acoustic Theory of Speech Production*. Walter de Gruyter, 1970. 344 pp. ISBN: 978-90-279-1600-6.

[63] F. Farace et al. "Free anterolateral thigh flap versus free forearm flap: Functional results in oral reconstruction." In: *Journal of Plastic, Reconstructive & Aesthetic Surgery* 60.6 (June 2007), pp. 583–587. ISSN: 17486815. DOI: 10.1016/j.bjps.2006.11.014. URL: https://linkinghub.elsevier.com/retrieve/pii/S1748681506006127 (visited on Sept. 10, 2022).

[64] S.-M. Fereshtehnejad et al. "Burden of neurodegenerative diseases in the Eastern Mediterranean Region, 1990–2016: findings from the Global Burden of Disease Study 2016." In: *European Journal of Neurology* 26.10 (Oct. 2019), pp. 1252–1265. ISSN: 1351-5101, 1468-1331. DOI: 10.1111/ene.13972. URL: https://onlinelibrary.wiley.com/doi/10.1111/ene.13972 (visited on June 3, 2022).

[65] John Field. "Psycholinguistics: The Key Concepts." In: (), p. 387.

[66] *Fine-Tune XLSR-Wav2Vec2 for low-resource ASR with 🤗 Transformers*. URL: https://huggingface.co/blog/fine-tune-xlsr-wav2vec2 (visited on Sept. 15, 2022).

[67] *FORCED ALIGNMENT ON CHILDREN SPEECH*. evalita. URL: https://www.evalita.it/campaigns/evalita-2014/tasks/forced-alignment-on-children-speech-facs/ (visited on Nov. 27, 2022).

[68] Mark J F Gales et al. "SPEECH RECOGNITION AND KEYWORD SPOTTING FOR LOW RESOURCE LANGUAGES: BABEL PROJECT RESEARCH AT CUED." In: (), p. 8.

[69] J. S. Garofolo et al. "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1." In: *NASA STI/Recon Technical Report N* 93 (Feb. 1, 1993). ADS Bibcode: 1993STIN...9327403G, p. 27403. URL: https://ui.adsabs.harvard.edu/abs/1993STIN...9327403G (visited on Nov. 28, 2022).

[70] *Gentle*. URL: https://lowerquality.com/gentle/ (visited on Nov. 26, 2022).

[71] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. *Unsupervised Representation Learning by Predicting Image Rotations*. Mar. 20, 2018. arXiv: 1803.07728[cs]. URL: http://arxiv.org/abs/1803.07728 (visited on Oct. 12, 2022).

[72] Christopher G. Goetz et al. "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results: MDS-UPDRS: Clinimetric Assessment." In: *Movement Disorders* 23.15 (Nov. 15, 2008), pp. 2129–2170. ISSN: 08853185. DOI: 10.1002/mds.22340. URL: https://onlinelibrary.wiley.com/doi/10.1002/mds.22340 (visited on June 6, 2022).

[73] Christopher G. Goetz et al. "*Movement* Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations The *Movement* Disorder Society Task Force on rating scales for Parkinson's disease." In: *Movement Disorders* 19.9 (Sept. 2004), pp. 1020–1028. ISSN: 08853185, 15318257. DOI: 10.1002/mds.20213. URL: https://onlinelibrary.wiley.com/doi/10.1002/mds.20213 (visited on Aug. 13, 2022).

[74] Kyle Gorman, Jonathan Howell, and Michael Wagner. "PROSODYLAB-ALIGNER: A TOOL FOR FORCED ALIGNMENT OF LABORATORY SPEECH." In: (), p. 3.

[75] Alex Graves et al. "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks." In: (2006), p. 8.

[76] Jordan R. Green et al. "Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases." In: *Interspeech 2021*. Interspeech 2021. ISCA, Aug. 30, 2021, pp. 4778–4782. DOI: 10.21437/Interspeech.2021-1384. URL: https://www.isca-speech.org/archive/interspeech_2021/green21_interspeech.html (visited on Aug. 23, 2022).

[77] Mónica Gómez-Benito et al. "Modeling Parkinson's Disease With the Alpha-Synuclein Protein." In: *Frontiers in Pharmacology* 11 (Apr. 23, 2020), p. 356. ISSN: 1663-9812. DOI: 10.3389/fphar.2020.00356. URL: https://www.frontiersin.org/article/10.3389/fphar.2020.00356/full (visited on June 5, 2022).

[78] J.A. Gómez-García, L. Moro-Velázquez, and J.I. Godino-Llorente. "On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art." In: *Biomedical Signal Processing and Control* 51 (May 2019), pp. 181–199. ISSN: 17468094. DOI: 10.1016/j.bspc.2018.12.024. URL: https://linkinghub.elsevier.com/retrieve/pii/S1746809418303239 (visited on June 12, 2022).

[79] J.A. Gómez-García, L. Moro-Velázquez, and J.I. Godino-Llorente. "On the design of automatic voice condition analysis systems. Part II: Review of speaker recognition techniques and study on the effects of different variability factors." In: *Biomedical Signal Processing and Control* 48 (Feb. 2019), pp. 128–143. ISSN: 17468094. DOI: 10.1016/j.bspc.2018.09.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1746809418302416 (visited on June 2, 2022).

[80] Robert A. Hall. "Italian Phonemes and Orthography." In: *Italica* 21.2 (1944). Publisher: American Association of Teachers of Italian, pp. 72–82. ISSN: 0021-3020. DOI: 10.2307/475860. URL: https://www.jstor.org/stable/475860 (visited on July 9, 2022).

[81] Awni Hannun. "Sequence Modeling with CTC." In: *Distill* 2.11 (Nov. 27, 2017), e8. ISSN: 2476-0757. DOI: 10.23915/distill.00008. URL: https://distill.pub/2017/ctc (visited on Sept. 15, 2022).

[82] Awni Hannun et al. *Deep Speech: Scaling up end-to-end speech recognition*. Dec. 19, 2014. arXiv: 1412.5567[cs]. URL: http://arxiv.org/abs/1412.5567 (visited on Sept. 1, 2022).

[83] Michael Henretty. *Sharing Our Common Voice — Mozilla Releases Second Largest Public Voice Data Set.* Mozilla Open Innovation. Nov. 29, 2017. URL: https://medium.com/mozilla-open-innovation/sharing-our-common-voice-mozilla-releases-second-largest-public-voice-data-set-e88f7d6b7666 (visited on Sept. 1, 2022).

[84] Moto Hira. *Forced Alignment with Wav2Vec2 — Torchaudio 0.12.1 documentation.* URL: https://pytorch.org/audio/0.12.1/tutorials/forced_alignment_tutorial.html (visited on Oct. 14, 2022).

[85] Jan Hlavnička et al. "Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder." In: *Scientific Reports* 7.1 (Dec. 2017), p. 12. ISSN: 2045-2322. DOI: 10.1038/s41598-017-00047-5. URL: http://www.nature.com/articles/s41598-017-00047-5 (visited on Aug. 16, 2022).

[86] Margaret M Hoehn. "Parkinsonism: onset, progression, and mortality." In: (), p. 17.

[87] *Home - Voiceitt.* URL: https://voiceitt.com/ (visited on Nov. 30, 2022).

[88] *How many languages are endangered?* Ethnologue. June 4, 2019. URL: https://www.ethnologue.com/guides/how-many-languages-endangered (visited on June 11, 2022).

[89] *How many languages are there in the world?* Ethnologue. May 3, 2016. URL: https://www.ethnologue.com/guides/how-many-languages (visited on June 11, 2022).

[90] Xuedong Huang, James Baker, and Raj Reddy. "A historical perspective of speech recognition." In: *Communications of the ACM* 57.1 (Jan. 2014), pp. 94–103. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/2500887. URL: https://dl.acm.org/doi/10.1145/2500887 (visited on Nov. 17, 2022).

[91] A J Hughes et al. "Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases." In: *Journal of Neurology, Neurosurgery & Psychiatry* 55.3 (Mar. 1, 1992), pp. 181–184. ISSN: 0022-3050. DOI: 10.1136/jnnp.55.3.181. URL: https://jnnp.bmj.com/lookup/doi/10.1136/jnnp.55.3.181 (visited on June 6, 2022).

[92] International Phonetic Association. *IPA Chart.* 2020. URL: https://www.internationalphoneticassociation.org/IPAcharts/IPA_chart_orig/pdfs/IPA_unitipa_2020_full.pdf (visited on July 10, 2022).

[93] Madeline Jefferson. "Usability of Automatic Speech Recognition Systems for Individuals with Speech Disorders:" in: (), p. 29.

[94] Zengrui Jin et al. *Adversarial Data Augmentation for Disordered Speech Recognition.* Aug. 2, 2021. arXiv: 2108.00899[cs,eess]. URL: http://arxiv.org/abs/2108.00899 (visited on Nov. 30, 2022).

[95] Keith Johnson. *Acoustic and auditory phonetics.* 2nd ed. Malden, Mass: Blackwell Pub, 2003. 182 pp. ISBN: 978-1-4051-0122-6 978-1-4051-0123-3.

[96] B H Juang and Lawrence R Rabiner. "Automatic Speech Recognition – A Brief History of the Technology Development." In: (), p. 24.

[97] Lisa Klingelhoefer and Heinz Reichmann. "Pathogenesis of Parkinson disease—the gut–brain axis and environmental factors." In: *Nature Reviews Neurology* 11.11 (Nov. 2015), pp. 625–636. ISSN: 1759-4758, 1759-4766. DOI: 10.1038/nrneurol.2015.197. URL: http://www.nature.com/articles/nrneurol.2015.197 (visited on June 4, 2022).

[98] Ludwig Kürzinger et al. "CTC-Segmentation of Large Corpora for German End-to-end Speech Recognition." In: *arXiv:2007.09127 [eess]* 12335 (2020), pp. 267–278. DOI: 10.1007/978-3-030-60276-5_27. arXiv: 2007.09127. URL: http://arxiv.org/abs/2007.09127 (visited on May 3, 2022).

[99] Peter Ladefoged. *Elements of acoustic phonetics*. 2nd ed. Chicago: University of Chicago Press, 1996. 216 pp. ISBN: 978-0-226-46763-4 978-0-226-46764-1.

[100] John Laver and Laver John. *Principles of Phonetics*. Cambridge University Press, May 12, 1994. 740 pp. ISBN: 978-0-521-45655-5.

[101] Juho Leinonen, Sami Virpioja, and Mikko Kurimo. "Grapheme-Based Cross-Language Forced Alignment: Results with Uralic Languages." In: (), p. 6.

[102] Federico Albano Leoni. *Il corpus CLIPS, presentazione del progetto a cura di Federico Albano Leoni*. 2007. URL: http://www.clips.unina.it/it/documenti/presentazione_clips.pdf (visited on Aug. 25, 2022).

[103] Vladimir I Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals." In: *Soviet physics doklady* 10.8 (1966), pp. 707–710. URL: https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf (visited on Sept. 11, 2022).

[104] Jingbei Li et al. *NeuFA: Neural Network Based End-to-End Forced Alignment with Bidirectional Attention Mechanism*. Mar. 31, 2022. arXiv: 2203.16838[cs,eess]. URL: http://arxiv.org/abs/2203.16838 (visited on Nov. 26, 2022).

[105] Jinyu Li. *Recent Advances in End-to-End Automatic Speech Recognition*. Feb. 2, 2022. arXiv: 2111.01690[cs,eess]. URL: http://arxiv.org/abs/2111.01690 (visited on Oct. 6, 2022).

[106] Bogdan Ludusan. "UNINA System for the EVALITA 2011 Forced Alignment Task." In: *Evaluation of Natural Language and Speech Tools for Italian*. Ed. by Bernardo Magnini et al. Red. by David Hutchison et al. Vol. 7689. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 330–337. ISBN: 978-3-642-35827-2 978-3-642-35828-9. DOI: 10.1007/978-3-642-35828-9_36. URL: http://link.springer.com/10.1007/978-3-642-35828-9_36 (visited on May 7, 2022).

[107] Fernando López and Jordi Luque. *Iterative pseudo-forced alignment by acoustic CTC loss for self-supervised ASR domain adaptation*. Oct. 27, 2022. arXiv: 2210.15226[cs,eess]. URL: http://arxiv.org/abs/2210.15226 (visited on Nov. 25, 2022).

[108] Andrew Ma, Kenneth K Lau, and Dominic Thyagarajan. "Voice changes in Parkinson's disease: What are they telling us?" In: *Journal of Clinical Neuroscience* 72 (Feb. 2020), pp. 1–7. ISSN: 09675868. DOI: 10.1016/j.jocn.2019.12.029. URL: https://linkinghub.elsevier.com/retrieve/pii/S0967586819317631 (visited on June 2, 2022).

[109] Robert L. MacDonald et al. "Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia." In: *Interspeech 2021*. Interspeech 2021. ISCA, Aug. 30, 2021, pp. 4833–4837. DOI: 10.21437/Interspeech.2021-697. URL: https://www.isca-speech.org/archive/interspeech_2021/macdonald21_interspeech.html (visited on Nov. 30, 2022).

[110] Luca Marsili, Giovanni Rizzo, and Carlo Colosimo. "Diagnostic Criteria for Parkinson's Disease: From James Parkinson to the Concept of Prodromal Disease." In: *Frontiers in Neurology* 9 (Mar. 23, 2018), p. 156. ISSN: 1664-2295. DOI: 10.3389/fneur.2018.00156. URL: http://journal.frontiersin.org/article/10.3389/fneur.2018.00156/full (visited on June 6, 2022).

[111] Israel Martínez-Nicolás et al. "Ten Years of Research on Automatic Voice and Speech Analysis of People With Alzheimer's Disease and Mild Cognitive Impairment: A Systematic Review Article." In: *Frontiers in Psychology* 12 (Mar. 23, 2021), p. 620251. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.620251. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.620251/full (visited on Aug. 21, 2022).

[112] A. Marzal and E. Vidal. "Computation of normalized edit distance and applications." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.9 (Sept. 1993), pp. 926–932. ISSN: 01628828. DOI: 10.1109/34.232078. URL: http://ieeexplore.ieee.org/document/232078/ (visited on Sept. 11, 2022).

[113] Pietro Maturi. *I suoni delle lingue, i suoni dell'italiano: introduzione alla fonetica*. Google-Books-ID: H5AqQgAACAAJ. Il Mulino, 2009. 159 pp. ISBN: 978-88-15-13305-2.

[114] Michael McAuliffe et al. "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." In: *Interspeech 2017*. Interspeech 2017. ISCA, Aug. 20, 2017, pp. 498–502. DOI: 10.21437/Interspeech.2017-1386. URL: https://www.isca-speech.org/archive/interspeech_2017/mcauliffe17_interspeech.html (visited on May 7, 2022).

[115] Michael McCloskey and Neal J. Cohen. "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem." In: *Psychology of Learning and Motivation*. Vol. 24. Elsevier, 1989, pp. 109–165. ISBN: 978-0-12-543324-2. DOI: 10.1016/S0079-7421(08)60536-8. URL: https://linkinghub.elsevier.com/retrieve/pii/S0079742108605368 (visited on Dec. 3, 2022).

[116] mmcauliffe. *Speaker dictionaries and multilingual IPA*. Michael McAuliffe. Section: Blog. Aug. 15, 2021. URL: https://memcauliffe.com/speaker-dictionaries-and-multilingual-ipa.html (visited on Sept. 11, 2022).

[117] mmcauliffe. *Update on Montreal Forced Aligner performance*. Michael McAuliffe. Section: Blog. Aug. 3, 2021. URL: https://memcauliffe.com/update-on-montreal-forced-aligner-performance.html (visited on May 7, 2022).

[118] Mahsa Mohaghegh and Jaya Gascon. "Identifying Parkinson's Disease using Multimodal Approach and Deep Learning." In: *2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA)*. 2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA). Nov. 2021, pp. 1–6. DOI: 10.1109/CITISIA53721.2021.9719945.

[119] Laureano Moro-Velazquez et al. "Advances in Parkinson's Disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects." In: *Biomedical Signal Processing and Control* 66 (Apr. 2021), p. 102418. ISSN: 17468094. DOI: 10.1016/j.bspc.2021.102418. URL: https://linkinghub.elsevier.com/retrieve/pii/S174680942100015X (visited on Aug. 23, 2022).

[120] Laureano Moro Velázquez. "Towards the differential evaluation of Parkinson's Disease by means of voice and speech processing." PhD thesis. Universidad Politécnica de Madrid, May 2018. DOI: 10.20868/UPM.thesis.51278. URL: http://oa.upm.es/51278/ (visited on May 14, 2022).

[121] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. "The Unified Parkinson's Disease Rating Scale (UPDRS): Status and recommendations." In: *Movement Disorders* 18.7 (July 2003), pp. 738–750. ISSN: 0885-3185, 1531-8257. DOI: 10.1002/mds.10473. URL: https://onlinelibrary.wiley.com/doi/10.1002/mds.10473 (visited on June 9, 2022).

[122] *Mozilla Common Voice*. URL: https://commonvoice.mozilla.org/ (visited on Sept. 1, 2022).

[123] *Mozilla Common Voice*. URL: https://commonvoice.mozilla.org/en/datasets (visited on Oct. 19, 2022).

[124] *Munich AUtomatic Segmentation*. URL: https://www.bas.uni-muenchen.de/Bas/BasMAUS.html (visited on Nov. 26, 2022).

[125] Quoc Cuong Ngo et al. "Computerized analysis of speech and voice for Parkinson's disease: A systematic review." In: *Computer Methods and Programs in Biomedicine* 226 (Nov. 2022), p. 107133. ISSN: 01692607. DOI: 10.1016/j.cmpb.2022.107133. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169260722005144 (visited on Nov. 10, 2022).

[126] Sieb Nooteboom. "THE PROSODY OF SPEECH: MELODY AND RHYTHM." In: (), p. 49.

[127] Juan Rafael Orozco-Arroyave. "Analysis of Speech of People with Parkinson's Disease." In: (), p. 138.

[128] *Orthophony*. In: *The Free Dictionary*. URL: https://www.thefreedictionary.com/Orthophony (visited on Aug. 31, 2022).

[129] *orthophony*. In: *Wiktionary*. Page Version ID: 62244056. Mar. 25, 2021. URL: https://en.wiktionary.org/w/index.php?title=orthophony&oldid=62244056 (visited on Aug. 31, 2022).

[130] Giulio Paci, Giacomo Sommavilla, and Piero Cosi. "SAD-Based Italian Forced Alignment Strategies." In: *Evaluation of Natural Language and Speech Tools for Italian*. Ed. by Bernardo Magnini et al. Red. by David Hutchison et al. Vol. 7689. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 322–329. ISBN: 978-3-642-35827-2 978-3-642-35828-9. DOI: 10.1007/978-3-642-35828-9_35. URL: http://link.springer.com/10.1007/978-3-642-35828-9_35 (visited on May 7, 2022).

[131] B Pakkenberg. "Aging and the human neocortex." In: *Experimental Gerontology* 38.1 (Jan. 2003), pp. 95–99. ISSN: 05315565. DOI: 10.1016/S0531-5565(02)00151-1. URL: https://linkinghub.elsevier.com/retrieve/pii/S0531556502001511 (visited on June 8, 2022).

[132] Vassil Panayotov et al. "Librispeech: An ASR corpus based on public domain audio books." In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2015 - 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5206–5210. ISBN: 978-1-4673-6997-8. DOI: 10.1109/ICASSP.2015.7178964. URL: http://ieeexplore.ieee.org/document/7178964/ (visited on Nov. 29, 2022).

[133] Daniel S. Park et al. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition." In: *Interspeech 2019*. Sept. 15, 2019, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680. arXiv: 1904.08779[cs,eess,stat]. URL: http://arxiv.org/abs/1904.08779 (visited on Nov. 30, 2022).

[134] James Parkinson. "An Essay on the Shaking Palsy." In: *J Neuropsychiatry Clin Neurosci* (2002), p. 14.

[135] Terence Parr. *The definitive ANTLR 4 reference*. The pragmatic programmers. OCLC: ocn802295434. Dallas, Texas: The Pragmatic Bookshelf, 2012. 305 pp. ISBN: 978-1-934356-99-9.

[136] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: (), p. 12.

[137] Alberto Pettarin. *forced-alignment-tools*. original-date: 2016-08-01T15:50:38Z. Nov. 26, 2022. URL: https://github.com/pettarin/forced-alignment-tools (visited on Nov. 26, 2022).

[138] Mark A. Pitt et al. "The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability." In: *Speech Communication* 45.1 (Jan. 2005), pp. 89–95. ISSN: 01676393. DOI: 10.1016/j.specom.2004.09.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167639304000974 (visited on Nov. 28, 2022).

[139] Werner Poewe et al. "Parkinson disease." In: *Nature Reviews Disease Primers* 3.1 (Dec. 21, 2017), p. 17013. ISSN: 2056-676X. DOI: 10.1038/nrdp.2017.13. URL: http://www.nature.com/articles/nrdp201713 (visited on June 3, 2022).

[140] R. B. Postuma et al. "How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder." In: *Brain* 135.6 (June 1, 2012), pp. 1860–1870. ISSN: 0006-8950, 1460-2156. DOI: 10.1093/brain/aws093. URL: https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/aws093 (visited on Aug. 13, 2022).

[141] Ronald B. Postuma et al. "MDS clinical diagnostic criteria for Parkinson's disease: MDS-PD Clinical Diagnostic Criteria." In: *Movement Disorders* 30.12 (Oct. 2015), pp. 1591–1601. ISSN: 08853185. DOI: 10.1002/mds.26424. URL: https://onlinelibrary.wiley.com/doi/10.1002/mds.26424 (visited on June 9, 2022).

[142] Daniel Povey et al. "The Kaldi Speech Recognition Toolkit." In: (), p. 4.

[143] Massimo Prada. "Breve introduzione alla fonetica." In: (), p. 134.

[144] Vineel Pratap et al. "MLS: A Large-Scale Multilingual Dataset for Speech Research." In: *Interspeech 2020*. Oct. 25, 2020, pp. 2757–2761. DOI: 10.21437/Interspeech.2020-2826. arXiv: 2012.03411[cs,eess]. URL: http://arxiv.org/abs/2012.03411 (visited on Oct. 11, 2022).

[145] *Project Euphonia*. URL: https://sites.research.google/euphonia/about/ (visited on Nov. 30, 2022).

[146] *Project Euphonia*. URL: https://sites.research.google/euphonia/faq/ (visited on Nov. 30, 2022).

[147] *prosody noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com*. URL: https://www.oxfordlearnersdictionaries.com/definition/english/prosody (visited on June 21, 2022).

[148] L.R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition." In: *Proceedings of the IEEE* 77.2 (Feb. 1989). Conference Name: Proceedings of the IEEE, pp. 257–286. ISSN: 1558-2256. DOI: 10.1109/5.18626.

[149] Alec Radford et al. "Robust Speech Recognition via Large-Scale Weak Supervision." In: (), p. 28.

[150] *Recreating Natural Voices for People with Speech Impairments*. URL: https://ai.googleblog.com/2021/08/recreating-natural-voices-for-people.html (visited on Nov. 30, 2022).

[151] Ingrid Rosenfelder et al. *FAVE: Forced alignment and vowel extraction*. Aug. 11, 2022. DOI: 10.5281/zenodo.7143517. URL: https://zenodo.org/record/7143517 (visited on Nov. 26, 2022).

[152] Anthony Rousseau, Paul Deléglise, and Yannick Estève. "Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks." In: (), p. 6.

[153] Nelson Roy et al. "Voice Disorders in the General Population: Prevalence, Risk Factors, and Occupational Impact:" in: *The Laryngoscope* 115.11 (Nov. 2005), pp. 1988–1995. ISSN: 0023-852X. DOI: 10.1097/01.mlg.0000179174.32345.41. URL: http://doi.wiley.com/10.1097/01.mlg.0000179174.32345.41 (visited on June 19, 2022).

[154] Robert J Ruben. "Redefining the Survival of the Fittest: Communication Disorders in the 21st Century." In: (2000), p. 5.

[155] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. "The TORGO database of acoustic and articulatory speech from speakers with dysarthria." In: *Language Resources and Evaluation* 46.4 (Dec. 2012), pp. 523–541. ISSN: 1574-020X, 1574-0218. DOI: 10.1007/s10579-011-9145-0. URL: http://link.springer.com/10.1007/s10579-011-9145-0 (visited on Nov. 30, 2022).

[156] J. Rusz et al. "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease." In: *The Journal of the Acoustical Society of America* 129.1 (Jan. 2011), pp. 350–367. ISSN: 0001-4966. DOI: 10.1121/1.3514381. URL: http://asa.scitation.org/doi/10.1121/1.3514381 (visited on June 8, 2022).

[157] Robert T. Sataloff, Yolanda D. Heman-Ackah, and Mary J. Hawkshaw. "Clinical Anatomy and Physiology of the Voice." In: *Otolaryngologic Clinics of North America* 40.5 (Oct. 2007), pp. 909–929. ISSN: 00306665. DOI: 10.1016/j.otc.2007.05.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0030666507000874 (visited on June 12, 2022).

[158] Renata Savy. *Specifiche per l'etichettatura dei livelli segmentali*. Mar. 10, 2006. URL: http://www.clips.unina.it/it/documenti/12_specifiche_di_etichettatura.pdf (visited on Aug. 26, 2022).

[159] Renata Savy and Valentina Caniparoli. *Descrizione dell'archivio di CLIPS*. 2007. URL: http://www.clips.unina.it/it/documenti/14_archivio_clips.zip (visited on Aug. 26, 2022).

[160] Benjamin G. Schultz et al. "Automatic speech recognition in neurodegenerative disease." In: *International Journal of Speech Technology* 24.3 (Sept. 2021), pp. 771–779. ISSN: 1381-2416, 1572-8110. DOI: 10.1007/s10772-021-09836-w. URL: https://link.springer.com/10.1007/s10772-021-09836-w (visited on Nov. 25, 2022).

[161] *SCTK, the NIST Scoring Toolkit*. original-date: 2016-05-03T19:00:47Z. Nov. 22, 2022. URL: https://github.com/usnistgov/SCTK/blob/f48376a203ab17f0d479995d87275db6772dcb4a/doc/sclite.htm (visited on Nov. 28, 2022).

[162] Joel Shor et al. "Personalizing ASR for Dysarthric and Accented Speech with Limited Data." In: *Interspeech 2019*. Sept. 15, 2019, pp. 784–788. DOI: 10.21437/Interspeech.2019-1427. arXiv: 1907.13511[cs,eess]. URL: http://arxiv.org/abs/1907.13511 (visited on Nov. 1, 2022).

[163] Alberto Sobrero and Immacolata Tempesta. *Definizione delle caratteristiche generali del corpus: informatori, località*. Mar. 10, 2006. URL: http://www.clips.unina.it/it/documenti/1_scelta_informatori_e_localita.pdf (visited on Aug. 26, 2022).

[164] Yuandong Tian, Xinlei Chen, and Surya Ganguli. *Understanding self-supervised Learning Dynamics without Contrastive Pairs*. Oct. 7, 2021. arXiv: 2102.06810[cs]. URL: http://arxiv.org/abs/2102.06810 (visited on Oct. 12, 2022).

[165] Ingo R. Titze. "Comments on the Myoelastic - Aerodynamic Theory of Phonation." In: *Journal of Speech, Language, and Hearing Research* 23.3 (Sept. 1980). Publisher: American Speech-Language-Hearing Association, pp. 495–510. DOI: 10.1044/jshr.2303.495. URL: https://pubs.asha.org/doi/10.1044/jshr.2303.495 (visited on July 3, 2022).

[166] Jimmy Tobin and Katrin Tomanek. *Personalized Automatic Speech Recognition Trained on Small Disordered Speech Datasets*. Oct. 9, 2021. arXiv: 2110.04612[cs, eess]. URL: http://arxiv.org/abs/2110.04612 (visited on Nov. 30, 2022).

[167] Isao Tokuda. "The Source–Filter Theory of Speech." In: *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, Nov. 29, 2021. ISBN: 978-0-19-938465-5. DOI: 10.1093/acrefore/9780199384655.013.894. URL: https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-894 (visited on July 5, 2022).

[168] *Using WaveNet technology to reunite speech-impaired users with their original voices*. URL: https://www.deepmind.com/blog/using-wavenet-technology-to-reunite-speech-impaired-users-with-their-original-voices (visited on Nov. 30, 2022).

[169] Jörgen Valk and Tanel Alumäe. *VoxLingua107: a Dataset for Spoken Language Recognition*. Nov. 25, 2020. arXiv: 2011.12998[eess]. URL: http://arxiv.org/abs/2011.12998 (visited on Oct. 11, 2022).

[170] den Berg Janwillem van. "Myoelastic-Aerodynamic Theory of Voice Production." In: *Journal of Speech and Hearing Research* 1.3 (Sept. 1958). Publisher: American Speech-Language-Hearing Association, pp. 227–244. DOI: 10.1044/jshr.0103.227. URL: https://pubs.asha.org/doi/abs/10.1044/jshr.0103.227 (visited on July 3, 2022).

[171] Ashish Vaswani et al. *Attention Is All You Need*. Dec. 5, 2017. arXiv: 1706.03762[cs]. URL: http://arxiv.org/abs/1706.03762 (visited on Oct. 4, 2022).

[172] Andrey Verendeev and Chet C Sherwood. "Human brain evolution." In: *Current Opinion in Behavioral Sciences* 16 (Aug. 2017), pp. 41–45. ISSN: 23521546. DOI: 10.1016/j.cobeha.2017.02.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352154616302327 (visited on June 8, 2022).

[173] Andrew J Viterbi. "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm." In: (), p. 10.

[174] *Voice Disorders*. American Speech-Language-Hearing Association. Publisher: American Speech-Language-Hearing Association. URL: https://www.asha.org/practice-portal/clinical-topics/voice-disorders/ (visited on June 19, 2022).

[175] Sarah Catherine Walpole et al. "The weight of nations: an estimation of adult human biomass." In: *BMC Public Health* 12.1 (Dec. 2012), p. 439. ISSN: 1471-2458. DOI: 10.1186/1471-2458-12-439. URL: http://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-12-439 (visited on June 8, 2022).

[176] Changhan Wang et al. "VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021, pp. 993–1003. DOI: 10.18653/v1/2021.acl-long.80. URL: https://aclanthology.org/2021.acl-long.80 (visited on Oct. 11, 2022).

[177] Chengyi Wang et al. *UniSpeech: Unified Speech Representation Learning with Labeled and Unlabeled Data*. June 10, 2021. arXiv: 2101.07597[cs,eess]. URL: http://arxiv.org/abs/2101.07597 (visited on Nov. 26, 2022).

[178] Dong Wang, Xiaodong Wang, and Shaohe Lv. "An Overview of End-to-End Automatic Speech Recognition." In: *Symmetry* 11.8 (Aug. 7, 2019), p. 1018. ISSN: 2073-8994. DOI: 10.3390/sym11081018. URL: https://www.mdpi.com/2073-8994/11/8/1018 (visited on Nov. 21, 2022).

[179] Song Wang and Guanyu Li. "Overview of end-to-end speech recognition." In: *Journal of Physics: Conference Series* 1187.5 (Apr. 2019), p. 052068. ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/1187/5/052068. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1187/5/052068 (visited on Oct. 6, 2022).

[180] Shinji Watanabe et al. *ESPnet: End-to-End Speech Processing Toolkit*. Mar. 30, 2018. arXiv: 1804.00015[cs]. URL: http://arxiv.org/abs/1804.00015 (visited on Nov. 29, 2022).

[181] Shinji Watanabe et al. "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition." In: *IEEE Journal of Selected Topics in Signal Processing* 11.8 (Dec. 2017), pp. 1240–1253. ISSN: 1932-4553, 1941-0484. DOI: 10.1109/JSTSP.2017.2763455. URL: http://ieeexplore.ieee.org/document/8068205/ (visited on May 3, 2022).

[182] J C Wells. "Computer-coding the IPA: a proposed extension of SAMPA." In: (), p. 18.

[183] *What is Few-Shot Learning? Methods & Applications*. URL: https://research.aimultiple.com/few-shot-learning/ (visited on Oct. 30, 2022).

[184] Sean White. *Announcing the Initial Release of Mozilla's Open Source Speech Recognition Model and Voice Dataset*. The Mozilla Blog. Nov. 29, 2017. URL: https://blog.mozilla.org/en/mozilla/announcing-the-initial-release-of-mozillas-open-source-speech-recognition-model-and-voice-dataset/ (visited on Sept. 1, 2022).

[185] *X-SAMPA*. In: *Wikipedia*. Page Version ID: 120318171. Apr. 30, 2021. URL: https://it.wikipedia.org/w/index.php?title=X-SAMPA&oldid=120318171 (visited on Sept. 26, 2022).

[186] Yao-Yuan Yang et al. *TorchAudio: Building Blocks for Audio and Speech Processing*. Feb. 16, 2022. arXiv: 2110.15018[cs,eess]. URL: http://arxiv.org/abs/2110.15018 (visited on Nov. 29, 2022).

[187] SJ Young. *The HTK Hidden Markov Model Toolkit: Design and Philosophy*. Sept. 6, 1994.

[188] Li Yujian and Liu Bo. "A Normalized Levenshtein Distance Metric." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.6 (June 2007), pp. 1091–1095. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.1078. URL: http://ieeexplore.ieee.org/document/4160958/ (visited on Sept. 11, 2022).

[189] Xiaohua Zhai et al. "S4L: Self-Supervised Semi-Supervised Learning." In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, Oct. 2019, pp. 1476–1485. ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00156. URL: https://ieeexplore.ieee.org/document/9010283/ (visited on Oct. 11, 2022).

[190] Zhaoyan Zhang. "Mechanics of human voice production and control." In: *The Journal of the Acoustical Society of America* 140.4 (Oct. 2016), pp. 2614–2635. ISSN: 0001-4966. DOI: 10.1121/1.4964509. URL: http://asa.scitation.org/doi/10.1121/1.4964509 (visited on June 25, 2022).

[191] Carl Zimmer. *100 Trillion Connections: New Efforts Probe and Map the Brain's Detailed Architecture*. Scientific American. DOI: 10.1038/scientificamerican0111-58. URL: https://www.scientificamerican.com/article/100-trillion-connections/ (visited on June 8, 2022).

[192] Jan G. Švec et al. "Integrative Insights into the Myoelastic-Aerodynamic Theory and Acoustics of Phonation. Scientific Tribute to Donald G. Miller." In: *Journal of Voice* (Mar. 2021), S0892199721000552. ISSN: 08921997. DOI: 10.1016/j.jvoice.2021.01.023. URL: https://linkinghub.elsevier.com/retrieve/pii/S0892199721000552 (visited on July 3, 2022).

Acronyms

AD Alzheimer's Disease. 11, 13, 40

ANN Artificial Neural Network. 41, 45, 102

ANTLR ANother Tool for Language Recognition. 86

ASR Automatic Speech Recognition. 3–5, 7, 12, 39, 42–52, 54, 57, 58, 65–68, 70, 72, 74, 75, 87, 89, 91, 92, 95, 101–103

AST Automatic Speech Translation. 74

AUC Area Under ROC Curve. 42

AVCA Automatic Voice Condition Analysis. 40

AVSCA Automatic Voice and Speech Condition Analysis. 12, 40, 42, 101–103

CER Character Error Rate. 48, 95

CLFA Cross-Language Forced Alignment. 51

CLIPS Corpora e Lessici di Italiano Parlato e Scritto. 3–5, 52–55, 57–62, 64, 65, 75, 76, 84–86, 89, 92, 95, 96, 101, 102

CMS Cepstral Mean Subtraction. 52

CTC Connectionist Temporal Classification. 3, 7, 47, 51, 67, 70, 71, 74, 75, 89–92

DALYs Disability-Adjusted Life Years. 13

DARPA Defense Advanced Research Projects Agency. 44

DET Detection Error Tradeoff. 42

DNN Deep Neural Network. 41, 46, 54

DTW Dynamic Time Warping. 51

E2E End-to-End. 3, 46, 47, 50, 51, 54, 55, 74, 101–103

F0 Fundamental frequency. 25, 26, 29, 37

F1 First formant. 29

F2 Second formant. 29

FA Forced Alignment. 3–5, 8, 11, 12, 39, 43, 49–55, 65, 66, 69, 70, 72, 74, 75, 87, 90–94, 96–98, 101–103

G2P Grapheme-to-Phoneme. 50, 54, 55, 95, 96, 102

GLD Generalized Levenshtein Distance. 67

GMM Gaussian Mixture Model. 45, 46, 50, 52–54

HMM Hidden Markov Model. 41, 45, 46, 50–53, 125

HTK Hidden Markov Model Toolkit. 50

HY Hoehn & Yahr. 2, 18, 65

IPA International Phonetic Alphabet. 4, 5, 7, 30, 31, 81–83, 85–87, 89, 102

IPVS Italian Parkinson's Voice and Speech. 3, 7, 43, 57, 65, 75, 89, 90, 95, 96, 98

k-NN k-Nearest Neighbors. 41, 43

LPC Linear Predictive Coding. 44

Lx Larynx waveform. 26

MAE Mean Absolute Error. 51

MAUS Munich AUtomatic Segmentation. 51

MDS Movement Disorder Society. 17, 18, 124

MDS-PD Movement Disorder Society (MDS) Clinical Diagnostic Criteria for Parkinson's Disease. 17

MDS-UPDRS Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale. 18

MFA Montreal Forced Aligner. 50, 51, 96

MIDA Mutual Information Discriminant Analysis. 52, 53

MLS Multilingual LibriSpeech. 95

MUR Ministero dell'università e della ricerca. 58

MURST Ministero dell'Università e della Ricerca Scientifica e Tecnologica. 58

ND Neurodegenerative Disease. 11, 13, 35, 40

NED Normalized Edit Distance. 68, 70

NINDS U.S. National Institute of Neurological Disorders and Stroke. 17

NIST National Institute of Standards and Technology. 53

NLP Natural Language Processing. 52, 72, 73

PBE Phone Boundary Error. 3, 66, 69, 75, 76, 92, 93, 96

PD Parkinson's Disease. 3, 11–19, 22, 35, 37–40, 42, 43, 57, 65, 83, 96, 98, 101–103

PER Phoneme Error Rate. 3, 66–68, 70, 74, 75, 90, 95

QALYs Quality-Adjusted Life Years. 13

RNN Recurrent Neural Network. 46, 47, 49, 70

ROC Receiver-Operating Curve. 42, 123

SAMPA Speech Assessment Methods Phonetic Alphabet. 63, 85

SCHMM Semi-Continuous HMM. 53

SCLITE Score Lite. 53

SCTK Scoring Toolkit. 53

SPRAAK Speech Processing, Recognition and Automatic Annotation Kit. 52, 53

STT Speech-to-Text. 44, 48

SUR Speech Understanding Research. 44

SVM Support Vector Machine. 41, 43

TR Transition Regions. 43

TTS Text To Speech. 51, 82

UKPDSBB United Kingdom Parkinson's Disease Society Brain Bank. 17

UPDRS Unified Parkinson's Disease Rating Scale. 2, 17, 18

VTLN Vocal Tract Length Normalization. 52

WBE Word Boundary Error. 3, 67, 70, 75, 76, 93, 96

WER Word Error Rate. 3, 48, 49, 66, 68, 69, 74, 75, 89, 90, 95, 96

X-SAMPA Extended Speech Assessment Methods Phonetic Alphabet. 4, 30, 62, 85, 86, 102

ZRSC Zero Resource Speech Challenge. 48