UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

DSIC
DEPARTAMENT DE SISTEMES
INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dept. of Computer Systems and Computation

Exploring multimodal foundation models to improve interaction for people with speech impairments.

Master's Thesis

Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

AUTHOR: Ferri Mollá, Isabel

Tutor: Linares Pellicer, Jordi Joan

ACADEMIC YEAR: 2022/2023

Resum

Les persones amb dificultats en la pronunciació, sovint derivades de patologies fisiològiques o cognitives, s'enfronten a reptes significatius en l'ús de tecnologies d'interacció per veu. Les tecnologies d'assistència actuals no aborden adequadament les complexitats úniques d'aquests reptes, destacant la necessitat de solucions adaptables per a millorar les capacitats de comunicació i la qualitat de vida.

Aquest projecte té com a objectiu adaptar diversos sistemes de reconeixement automàtic de la parla a un grup demogràfic específic, particularment a individus amb problemes de pronunciació, especialment aquells amb afàsia. Per a aconseguir-ho, s'aplicarà un procés de fine-tuning a diferents sistemes de reconeixement de la parla preentrenats, amb èmfasi en la identificació d'hiperparàmetres òptims per a l' entrenament així com en la comparació de resultats utilitzant la mètrica del Word Error Rate (WER).

A més, el projecte integrarà models de reconeixement de la parla amb models de descripció d'imatges per explorar en quina mesura el context visual millora la interpretació i comprensió del sistema sobre el que els individus amb afàsia estan intentant comunicar. L'avaluació d'aquests sistemes inclourà avaluacions humanes. Així, aquest projecte busca crear una solució integral per a ajudar les persones amb afàsia i millorar l'experiència d'Interacció Persona-Ordinador (HCI) per a aquest grup demogràfic.

Paraules clau: HCI; model de lenguatge; ASR; problemes de dicció; descripció d' imatges; foundation models.

Resumen

Las personas con dificultades en la pronunciación, a menudo derivadas de patologías fisiológicas o cognitivas, enfrentan desafíos significativos al utilizar tecnologías de interacción por voz. Las tecnologías de asistencia actuales no abordan adecuadamente las complejidades únicas de estos desafíos, lo que destaca la necesidad de soluciones adaptables para mejorar las capacidades de comunicación y la calidad de vida.

Este proyecto tiene como objetivo adaptar varios sistemas de reconocimiento automático del habla a un grupo demográfico específico, en particular, a individuos con problemas de pronunciación, especialmente aquellos con afasia. Para lograrlo, se realizará un proceso de fine-tuning a diferentes sistemas de reconocimiento del habla preentrenados, con énfasis en la identificación de hiperparámetros óptimos para el entrenamiento y en la comparación de resultados utilizando la métrica del Word Error Rate (WER).

Además, el proyecto integrará modelos de reconocimiento del habla con modelos de descripción de imágenes para explorar en qué medida el contexto visual mejora la interpretación y comprensión del sistema sobre lo que los individuos con afasia están tratando de comunicar. La evaluación de estos sistemas incluirá valoraciones humanas. Así, este proyecto busca crear una solución integral para ayudar a las personas con afasia y mejorar la experiencia de Interacción Persona-Ordenador (HCI) para este grupo demográfico.

Palabras clave: HCI; modelo de lenguaje; ASR; problemas de dicción; descripción de imágenes; foundation models.

Abstract

People with pronunciation difficulties, often stemming from physiological or cognitive pathologies, face significant challenges when using voice interaction technologies. Current assistive technologies do not adequately address the unique complexities of these challenges, highlighting the need for adaptable solutions to enhance communication abilities and quality of life.

This project aims to adapt various automatic speech recognition systems to a specific demographic group, particularly individuals with pronunciation problems, especially those with aphasia. To achieve this, fine-tuning will be applied to different pre-trained speech recognition systems, with a focus on identifying optimal hyperparameters for training and comparing results using the Word Error Rate (WER) metric.

Furthermore, the project will integrate speech recognition models with image description models to explore to what extent visual context enhances the system's interpretation and understanding of what individuals with aphasia are trying to communicate. Evaluation of these systems will include human assessments. Thus, this project seeks to create a comprehensive solution to assist people with aphasia and enhance the Human-Computer Interaction (HCI) experience for this demographic group.

Key words: HCI; language model; ASR; pronunciation problems; image captioning; foundation models

Contents

Appendix

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation

Speech recognition is the task of detecting spoken language and transcribing it into written text. Spoken language constitutes one of the primary avenues of communication used by a substantial majority of the human population to express themselves and exchange ideas. This circumstance, in conjunction with the remarkable progress experienced by technology in recent years, has led to a significant increase in the investigation and advancement within the field of Automatic Speech Recognition (ASR).

ASR systems have demonstrated their significant utility across diverse domains, including chatbots, voice assistants, and translation systems. The advancement in this field in the recent years, has resulted in the coexistence of current systems ranging from classical solutions based on Hidden Markov Models and Natural Language Processing, like the Kaldi project [43], to more contemporary approaches that harness deep learning techniques. Between the systems that use deep learning to address this challenge it is possible to find open-source solutions, such as whisper [45], wav2vec 2.0 [7], and OpenSeq2Seq [29] as well as commercial solutions like Google Cloud Speech-to-Text and Microsoft Azure Speech Services .

The recent expansion of ASR technology has allowed its use as an additional mode of interaction with a wide range of systems and devices, further emphasizing the importance of making this technology more accessible to all types of users.

However, there is a percentage of the population, specifically more than 1.4% of the global population[8], that suffers from some form of speech disorder, making challenging for them to use speech in a conventional manner to communicate. These difficulties can lead to the fact that, despite the impressive performance demonstrated by cutting-edge systems of this kind for the majority of the population, current voice recognition systems face challenges when transcribing the speech of specific user groups with pronunciation difficulties. This, among other things, generates frustration among these communities and restricts their use of these technologies.

Despite advances and research in the field [5, 53], the complexity and variations in pronunciation continue to pose obstacles to developing an open-access model that guarantees a satisfactory transcription experience for people with speech disabilities.

However, fine-tuning of the most recent models, equipped with millions of parameters, for specific domains, is yielding highly promising results in terms of adapting systems to particular user groups. Examples of this phenomenon are evident in [50, 28]. These studies not only invite optimism regarding the possibility of adapting speech recognition systems to specific groups of individuals with pronunciation difficulties but also suggest the potential development of systems that assist in the communication of these individuals with pronunciation problems with the rest of the population. That is why this project will conduct experiments and study the effectiveness of systems that would not only transcribe what people with pronunciation difficulties are saying but, with the help of image description systems, would extract the user's visual context and use it to interpret what the user is trying to convey with the spoken phrase. This could serve as assistance for people with communication difficulties to have effective communication with a wider range of individuals in the general population.

In this project, experiments combining ASR and image description systems using LLM (Large Language Models) will be conducted, and it will be investigated whether this combination of systems offers advantages compared to simply adapting ASR systems.

1.2 Objectives

The primary objective of this project is to enhance the interaction between individuals with pronunciation difficulties and current speech recognition systems, aiming to improve these systems' understanding of people facing such challenges.

Specifically, we will compare and seek to enhance the metrics obtained using base systems adjusted with the AphasiaBank corpus [38], which contains recordings and transcriptions of individuals with aphasia. To achieve this, we will select different models and perform fine-tuning on them. Additionally, we will also investigate whether the visual context information of the individual can help better interpret what they intend to convey with the spoken phrase, potentially improving communication for these individuals with a broader segment of society. To do this, we will employ image description systems to leverage the visual information that can be obtained from the participant's environment. This information, along with the transcriptions from the fine-tuned model, will be used as input for a large language model. The objective of this multimodal system is to generate a message that aligns more closely with the speaker's intention.

To accomplish these primary objectives, we will pursue the following sub-goals:

1. Filtering and collecting the corpus data, including transcriptions and videos of individuals with Aphasia, stored within the database.

2. Data cleaning that involves adjusting the transcriptions to conform to the input format required by each model and segmenting both audio and transcriptions into smaller fragments.

3. Partition the data into distinct subsets: training, test, and validation, and engage in a discussion about determining the most suitable partitioning strategy.

4. Selecting appropriate ASR models and adjusting model weights to align with transcriptions from individuals encountering challenges in speech articulation.

5. Identify and choose a pertinent system with the capability to generate image descriptions.

6. Examine whether enhancing speech with contextual image descriptions, which depict what the subject might be observing, enhances the overall comprehension of speech for individuals with pronunciation difficulties.

1.3 Paper Structure

The following document has been divided into various chapters and sections, so that the necessary information to comprehend the experiments conducted and to achieve the defined objectives, in addition to the work carried out in this paper, is conveyed efficiently. Below, the different chapters will be described.

In Chapter 1, the motivation, objectives of the study, and the document's structure are explained.

Chapter 2 discusses various language disorders, including aphasia, the type of disorder that will be addressed in this study.

In Chapter 3, a review of the state-of-the-art in speech recognition systems will be conducted, with special emphasis on systems designed for people with disabilities or those that utilize multimodal approaches to complement transcription.

Subsequently, Chapter 4 will focus on Automatic Speech Recognition (ASR) systems. It will delve into their evolution, functioning, different approaches, and types of systems.

Chapter 5 covers different ASR systems currently in use, along with an explanation of the fine-tuning process for such systems.

Chapter 6 will explore image captioning systems and their operation. Additionally, it will describe some of the most relevant image captioning and Visual Transformer systems in use today.

As for Chapter 7, it will discuss the evolution and current usage of language models.

Chapter 8 will describe the solution proposed by this study, as well as the dataset that will be used.

In Chapter 9, the conducted experiments will be elucidated, along with the evaluation results of the various systems.

Finally, in Chapter 10, a summary of the key points, conclusions, and future work will be presented.

CHAPTER 2

Language disorders

2.1 Types of language disorders

Language disorders [49] are communication problems that can manifest as difficulties in language comprehension or expression, and they can affect both spoken and written language. Below, we will list and briefly explain some of the major language disorders.

Stuttering, commonly known as dysphemia, is a speech disorder in which the individual experiences disruptions that interrupt the normal flow of speech. In addition to repetition, it is possible that in their speech, sounds, syllables, or words last longer than usual. All of these factors together can lead to a lack of fluency.

On the other hand, there is the mixed receptive-expressive language disorder, which hinders individuals in their understanding and expression of language. Examples of this include using simple sentences, having a limited vocabulary, or experiencing comprehension difficulties.

Dyslexia is another disorder that involves difficulties in reading, stemming from issues with identifying speech sounds and relating them to letters and words.

Another common problem may be dysarthria. This condition involves muscle-related issues that interfere with speech production, making it difficult for those with this condition to articulate words clearly.

Furthermore, there is dysphasia, or specific language impairment, which results from a lack of coordination in words due to a cerebral injury.

Finally, there is aphasia, a problem caused by damage to the parts of the brain responsible for language. People suffering from aphasia may also experience other issues, such as dysarthria or apraxia. Since aphasia is the problem that will be predominantly discussed in this article, it will be addressed in greater detail in the following section.

2.2 Aphasia

As mentioned earlier, aphasia [16] is a disorder that affects both language comprehension and expression and is caused by dysfunctions in specific regions of the brain. The condition manifests as a failure in the bidirectional translation that establishes the correspondence between thoughts and language. In individuals with aphasia, the ability to accurately convert sequences of non-verbal mental representations that constitute thought into the symbols and grammatical structures of language is compromised. In other words, images or representations of thought can no longer be adequately expressed in words and phrases. Likewise, the reverse process, which involves generating mental images that correspond to a sentence heard or read, can also be affected in aphasia.

This problem is not limited to languages based on auditory signals but also affects sign languages. Additionally, it can affect the written code of any type of language, whether auditory-based like English or ideogram-based like several Asian languages. Aphasia can compromise multiple aspects of language, including syntax (the grammatical structure of sentences), lexicon (the set of words denoting meanings), and word morphology (the combination of individual speech sounds, known as phonemes, into smaller meaningful units called morphemes). In many cases, more than one of these aspects is affected in the same patient. For example, a patient may primarily have difficulties in sentence comprehension, along with a moderate inability to select the right word to express their thoughts. In another patient, the difficulty may lie in the inability to construct the appropriate grammatical structure for the thoughts they are trying to express.

Regarding the types of aphasia, this classification is based on the affected area of the brain and the symptoms that manifest. Here, some of the main types are described:

- Global Aphasia: This is the most severe type of aphasia and is applied to patients with significant difficulties in understanding spoken language, or they do not understand it at all. They may also pronounce few recognizable words and are unable to read or write.

- Mixed Non-Fluent Aphasia: In this type of aphasia, patients have limited speech and struggle with oral production. Additionally, their oral comprehension is limited, and they have a very basic level of reading and writing.

- Broca's Aphasia: Broca's aphasia is characterized by reasonable oral comprehension but significant issues with speech fluency. Individuals with this condition tend to use very short expressions.

- Wernicke's Aphasia: In this type of aphasia, the ability to comprehend the meaning of spoken words is primarily impaired. Although it generally does not affect speech fluency, the sentences produced often lack coherence. Additionally, reading and writing are typically severely affected.

- Anomic Aphasia: Anomic aphasia refers to the patients' inability to find the right words to express what they want to communicate, especially important nouns and verbs. While their speech is usually grammatically fluid, they experience difficulty and frustration in finding the right words. Typically, they have a good understanding of spoken language and adequate reading skills.

A more visual classification of the types of aphasia can be seen in the figure 2.1.

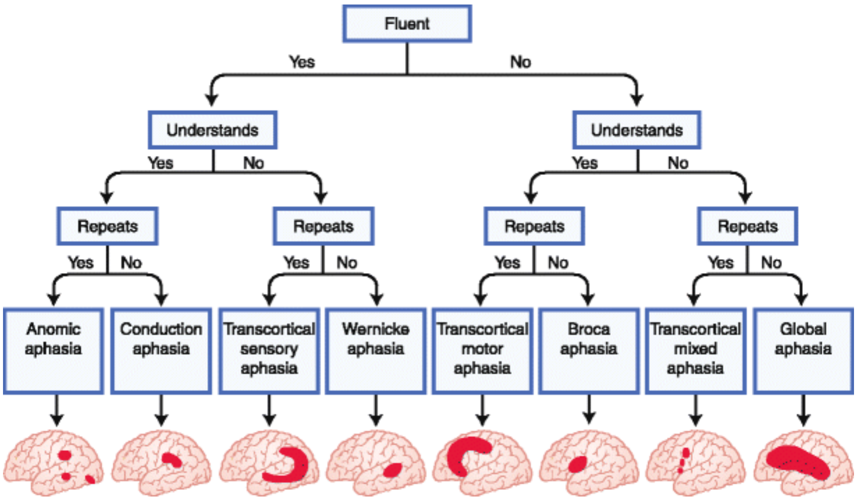


Figure 2.1: Decision-making tree in aphasic patients [11].

CHAPTER 3
Related Work

Since the shift in the ASR field from purely statistical models to incorporating neural networks, there has been a significant boom. The most significant change was the replacement of Gaussian mixtures in the acoustic model with deep neural networks [32]. However, the current state-of-the-art lies in end-to-end models. These models have revolutionised the field by employing a single neural network for transcribing audio signals to text.

As will be elucidated in Chapter 4, there are several advantages to using end-to-end models in the field of ASR. Some of these end-to-end advantages include the fact that they have a single objective function to optimize. Additionally, having fewer intermediate components makes the end-to-end solution more compact.

Automatic Speech Recognition is a challenging task, and its accuracy can be influenced by various factors such as ambient noise, speaker characteristics, or the usage of domain-specific vocabulary. These variations can significantly impact the system's performance.

Fine-tuning is a strategy that leverages pre-trained models and tailors them to a narrower, more specific domain. During this process, the model's weights are adjusted. This technique comes with several advantages. Since the pre-trained model already possesses prior knowledge, instead of training an entirely new model, this existing knowledge is harnessed. As a result, adapting a model, rather than starting from scratch, brings about a reduction in required training time and the volume of necessary data. This fine-tuning technique empowers models to adjust to specialized domains, delivering solutions that are more focused and efficacious.

Regarding fine-tuning focused on adapting speech recognition models for individuals with pronunciation difficulties, there have been several relevant research efforts.

In the research study by Green et al. (2021) [22], speech recognition models were tailored to individuals with speech disorders. Specifically, they evaluated speaker-independent models, such as Google's Speech-to-Text API, or end-to-end models based on RNN-T architecture [47]. The performance of each model was investigated, along with the improvements achieved by personalising the latter model to suit a specific participant with a speech disorder. Notably, through this adaptation process, significant word error rate reductions of up to 80% were observed in individuals with moderate to severe speech disorders. However, it is important to highlight that, in this particular study, a distinct and personalised model was developed for each speaker. In [50], several experiments finetuning ASR model for people with amyotrophic lateral sclerosis (ALS) and accented speech being able to reduce WER rater by up to 62% in the case of ALS.

Furthermore, speech is a dynamic and continuous process that is influenced not only by the linguistic content it conveys but also by the context in which the message is delivered. Considering the possibility of integrating other systems, such as image description, to enhance comprehension is a noteworthy aspect to explore.

Several comprehensive studies have been conducted in this domain. Notably, in [30] an innovative approach that utilizes image captions generated by a system captioning system as prompts for the task of correcting output texts from Automatic Speech Recognition systems is proposed. The authors introduced two distinct methods to accomplish this objective.

The first method, termed the "GatedFusion-based method," involves concatenating the embeddings of image visual features and the ASR transcription text. Subsequently, this combined information is passed through a decoder to rectify any transcription errors by leveraging the accompanying visual information.

The second approach, referred to as the "Prompt-based method", departs from embedding the image caption and ASR transcription separately. Instead, both the image caption and the ASR transcription are employed as prompts to an Encoder-Decoder model. Through this method, the image caption can effectively correct any inaccuracies present in the ASR transcription. It is imperative to acknowledge that the study, while undoubtedly valuable in its scope, focuses exclusively on general spoken English audios and their corresponding transcriptions. As such, the research investigates and evaluates the proposed methods based on this particular dataset. Nevertheless, it is essential to recognize that the study's findings and conclusions might not be directly representative for individuals with speech impairments.

In [18], a comprehensive pipeline that addresses various tasks, including speech-to-text, text-to-speech, image-to-text, and text-to-image. While this pipeline offers a holistic approach to tackle these tasks, it is particularly noteworthy for its potential applicability in ASR tasks.

Even in scenarios where voice and text data resources become limited or unavailable, there remains a promising avenue to enhance ASR performance by leveraging multimodal data and exploiting the information provided within the data chain. By incorporating diverse sources of data, such as images and their associated text, ASR systems can benefit from the complementary nature of multimodal information, leading to improved accuracy and robustness.

On the other hand, a error correction model is used and to obtain the final transcription, a n-best list of error correction model output is rescored with language model and visual semantic joint embedding, so visual information is used too to improve the speech recognition results.

In contrast, in in [13] an error correction model to enhance the performance of speech recognition systems is introduced. Their approach involves generating an n-best list of outputs using the error correction model. Subsequently, to obtain the final transcription, this n-best list is rescored utilizing both a language model and a visual semantic joint embedding technique. The inclusion of visual information in this process serves as a means to further improve the accuracy and efficacy of the speech recognition results.

In the case of the aforementioned multimodal systems, the inclusion of visual information primarily aims to rectify potential transcription errors. However, within the specific context of this work, the objective extends beyond that. When developing systems tailored for individuals with aphasia, in this instance, the goal is not solely to achieve the most accurate transcription possible. With multimodal systems, where the visual context of an image is incorporated, the aim is also to effectively comprehend what the user is attempting to communicate.

CHAPTER 4

Automatic Speech Recognition

4.1 ASR evolution

The origins of Automatic Speech Recognition (ASR) can be traced back to around 1952 when initial research and projects began in this field. During that time, the technology developed by Bell Laboratories in the context of Audrey program showed promising capabilities in recognizing numerical digits. In the subsequent years, research and work in speech recognition continued to evolve, and it was in the early 1970s that statistical models were first applied, thanks to the pioneering efforts of Frederick Jelinek.

However, it was not until the 1980s, when Jack Ferguson introduced the popular use of Hidden Markov Models (HMMs), that a significant shift occurred in ASR. This marked a departure from simple pattern recognition methods, which were based on templates and spectral distance measures, to a more sophisticated statistical approach for speech processing. HMMs became the state-of-the-art technique in ASR for many years, gaining popularity due to their standardized benchmarking capabilities and their ability to provide reliable results, as explained in [25].

Furthermore, the availability of large speech corpora for researchers to use in training and testing models greatly contributed to the advancement of ASR during this period. Towards the late 1980s, the integration of neural networks with HMMs emerged as a promising avenue to enhance existing models. This combination allowed for improved phoneme differentiation and the generation of coherent textual phrases.

In recent years, particularly since 2014, there has been a predominant advocacy for fully neural approaches with end-to-end systems in the field of Automatic Speech Recognition. This shift is attributed to significant technological advancements, the availability of large datasets, and the emergence of powerful processing units such as GPUs (Graphics Processing Units). These neural methods have shown remarkable improvements in accuracy and performance, particularly as they benefit from an increasing abundance of data.

The adoption of end-to-end systems in ASR has streamlined the overall process, as it allows for direct mapping from input audio to output text without relying on intermediate stages or handcrafted features. Neural networks, especially deep learning models, have demonstrated their prowess in learning complex patterns and representations from vast amounts of data, leading to enhanced recognition capabilities. The integration of end-to-end systems with neural networks has facilitated substantial enhancements in the performance of statistical models, thus shaping the current state-of-the-art in ASR as can be seen in [26].

4.2 Traditional ASR process

A classical Automatic Speech Recognition (ASR) system encompasses several pivotal stages for the transformation of audio signals into transcriptions. Among the key phases of the system are feature extraction, acoustic modeling, language modeling, and decoding [39].

Regarding the feature extraction it is the process of extracting features from the audio recordings. A feature identify specific characteristics of voice such as volume, accent, pitch.

The audio signal is pre-treated, including noise reduction, or discretized to a specific frequency (typically 16 kHz, but 8 kHz for telephone calls). The signal is then divided into frames, to which mathematical transformations such as the Fourier transform are applied to obtain representative coefficients.

Subsequently, after the feature extraction process, it is the turn of the acoustic model, which is trained on large volumes of data to enable it to recognise the relationship between acoustic features and the corresponding phonetic units or sub-word representations. The model captures and learns the associations between acoustic characteristics and linguistic units. Regarding the models themselves, they can be based on traditional techniques such as Markov models or on neural networks such as DNNs or CNNs.

In relation to language models, they are responsible for learning and understanding the grammar, structure, and context of language, while the acoustic model focuses on the sounds and pronunciation of speech. Concerning the language model, its task is to estimate the probability of a sequence of words occurring together, thus aiding in improving the precision of the system's transcription. In the past, classic approaches used N-gram models, whereas more recently, neural network-based approaches like RNN or Transformer have been employed.

At this point, the ASR system needs to decode the audio by combining the outputs of both models, the acoustic model and the language model, to obtain the most probable transcription of the input audio. Algorithms such as beam search or dynamic type wrapping are used for this purpose, aligning the outputs of the acoustic and language models to generate the transcription.

After the decoding process, the generated transcription may undergo post-processing techniques, which involve the application of additional Natural Language Processing (NLP) methods to enhance the quality of the transcription. These techniques encompass various tasks, such as error or grammar correction, sentiment analysis, punctuation insertion, and context-based filtering [41].

Although traditionally, in the field of automatic speech recognition, hybrid models consisting of previously explained components have been used, recently, end-to-end systems have been gaining popularity. Regarding hybrid systems, their components were trained separately and then connected to form the model. However, this methodology had certain disadvantages, such as the need to adapt different components and technologies within the hybrid model, as well as the creation of lexicons that mapped phonemes to words. Additionally, a critical aspect was the alignment of the acoustic model, which could introduce errors during acoustic training.

To address these limitations, a solution emerged in the form of end-to-end models [44]. These models are composed of a single unitary block that simultaneously optimizes all components, namely, the acoustic model and the language model, during the training process.

This evolution in the methodology of automatic speech recognition has proven to be promising by overcoming the inherent drawbacks of hybrid models. By unifying and optimizing all components in a single training process, end-to-end models have improved the efficiency and accuracy of automatic speech recognition, thereby reducing errors caused by acoustic model alignment and eliminating the need to create phonetic lexicons.

Furthermore, research also delves into how these models enhance speed and efficiency. Co-training end-to-end models for speech recognition and endpointing reduces latency, as demonstrated in Bijwadia's work [9]. This reinforces the notion that fundamentally, the perspective of discarding manual components and concentrating on holistic optimization underscores the potential of end-to-end models for the evolution of speech recognition. In addition to this research show that these models excel in addressing challenges such as noisy or accented speech by directly optimizing the entire system. An example of this is the study conducted by Kim in 2020 [27], which highlights the precision of the end-to-end model in recognizing speech from diverse groups, including children, the elderly, and those with non-conventional pronunciation.

4.3 Approaches in Automatic Speech Recognition

Among the most popular end-to-end approaches in ASR [33], is Connectionist Temporal Classification (CTC) [21], which maps the input speech sequence to an output sequence of words. In such systems, as the length of the output labels is shorter than that of the input sequence, blank labels are inserted between each output label to construct CTC paths of the same length as the input sequence.

Another commonly used ASR approach is the Attention-based Encoder-Decoder (AED) model [12], which is considered a leading solution in the field of end-to-end models, known for its effectiveness. The AED model consists of an encoder, an attention layer, and a decoder. It operates by probabilistically generating output sequences by taking into account previous decoder outputs and the original label sequence. In the training of the AED model, the goal is to minimize the negative probability associated with the correct labels, given the input speech sequence.

In a manner akin to the CTC methodology, the encoder facilitates the transformation of input features into concealed representations. Meanwhile, the attention module undertakes the calculation of cross-attention weights. These weights establish a linkage between the decoder's previous output and the encoder's output. Through this mechanism, contextual embeddings and labels obtained within the decoder are harnessed by the encoder to iteratively generate output. Notably, this iterative approach averts any assumptions of conditional independence.

To foster a more precise alignment between spoken signals and their corresponding labels, numerous AED models embrace a multi-task learning paradigm. In this scheme, they undergo optimization alongside a language model, a synergy achieved by sharing the encoder. However, it is prudent to acknowledge that while AED models excel in ASR translation, the unbridled application of attention across the entire input can potentially introduce performance detriments due to notable latency. Consequently, certain AED systems judiciously restrict attention to select segments of the input signal, a demarcation achieved through various delineation strategies. While using this latter approach allows AED to be used for real-time transcription, it still faces challenges regarding latency. For this reason, RNN Transducer is commonly used for end-to-end speech recognition to address latency issues.

RNN Transducer models are widely employed in real-time ASR. In this context, their output relies on the preceding token outputs and the speech sequence up to the current time.

The RNN-T (Recurrent Neural Network Transducer) [20] is composed of three key elements: the encoder, the prediction network, and the joint network.

The encoder processes input feature sequences and generates a high-level representation. The prediction network uses the model's previous output to produce another high-level representation. Subsequently, the joint network combines both representations to compute the probability of each output token.

Unlike the CTC model, the RNN-T eliminates the assumption of conditional independence, leading to enhanced performance. The loss function involves the logarithm of the sum of all possible alignments mapped to the label sequence.

To achieve low latency, various strategies have been employed, including the restricted alignment approach. This technique restricts the alignment during training within a real-time delay threshold, resulting in GPU memory savings and accelerated training. Although other strategies also aim to reduce latency, they may entail a trade-off with precision.

4.4 ASR Systems

Based on the level of training required for each system, they can be categorised into three distinct types: speaker dependent, speaker independent, and speaker adaptable.

Speaker-dependent systems necessitate prior specific training before usage, enabling them to adapt to the specific characteristics of the user utilising the model. These systems generally perform admirably for the individual on whom their training has focused, but their accuracy significantly diminishes when faced with a new speaker.

On the contrary, speaker-independent systems do not require prior specific training to comprehend and adapt to the speaker. The data used to train such systems typically encompass a vast number of speakers, allowing them to recognise different speakers accurately even if those individuals were not explicitly included in the training set. While these systems do not require prior training and achieve acceptable accuracy, they may exhibit lower precision when transcribing speech from a particular speaker, as compared to the speaker-dependent counterparts.

Speaker-adaptive systems, akin to speaker-independent ones, do not necessitate prior training for initial usage. However, they continuously adapt to the speaker's characteristics throughout their use.

Furthermore,as it can be seen in [60], ASR systems are further classified based on the input they receive, falling into categories such as isolated/discrete-word recognition, connected word recognition, and continuous speech recognition.

CHAPTER 5
Use of ASR systems

In this chapter, several of the ASR models currently accessible will be explored. These models will be chosen for the purpose of finetuning and tailoring them to the specific problem at hand, as part of the adaptation process.

5.1 Model Discussion

There are many speech recognition models available nowadays; specifically for this project, we will be focusing on end-to-end systems.

As previously discussed in the section 4.2 end-to-end speech recognition systems offer distinct advantages backed by various studies. Hence, in this project, the discussion pertaining to the ASR model will predominantly revolve around end-to-end systems. Between the different ASR models available several models are going to be finetuned with the data obtained from Aphasia Bank after a data preprocessing. Concretelly the different experiments will be carried out using whisper[45] as well as wav2vec 2.0[7]. This models are some of the most popular models in the hugging-face leaderboard.

Several other models also hold promise, such as ESPnet [58], an ASR system written in Python using PyTorch. It follows a similar approach to Kaldi [43] for data processing and is designed to be multilingual. Initially, this system was primarily focused on speech recognition; however, starting from 2020, this framework has expanded its capabilities to encompass text-to-speech, voice conversion, speech translation, speech enhancement, beamforming, speech separation, denoising, and dereverberation.

In addition to these, the most widely used current ASR models include Vosk [2], a system developed in 2020 with support for 10 languages and 50 MB portable models, and SpeechBrain [48], which was launched in 2021. SpeechBrain is a PyTorch-based transcription toolkit that supports various functions, such as speech-to-text, speaker verification, speaker diarization, speech separation, and speech enhancement, among others. Next, we will delve more deeply into the structure of the two models that will be chosen for fine-tuning.

5.1.1. Whisper

Whisper[45] is a pre-trained model for user speech recognition developed by OpenAI. It has been trained on 680,000 hours of multilingual and multitask supervised data gathered from the internet. This model enables the transcription of audio in multiple languages into text, as well as translation.

In terms of its architecture, Whisper adheres to a classical autoregressive encoder-decoder structure, resembling the original Transformer architecture[56]. This structure can be observed in Figure 5.1

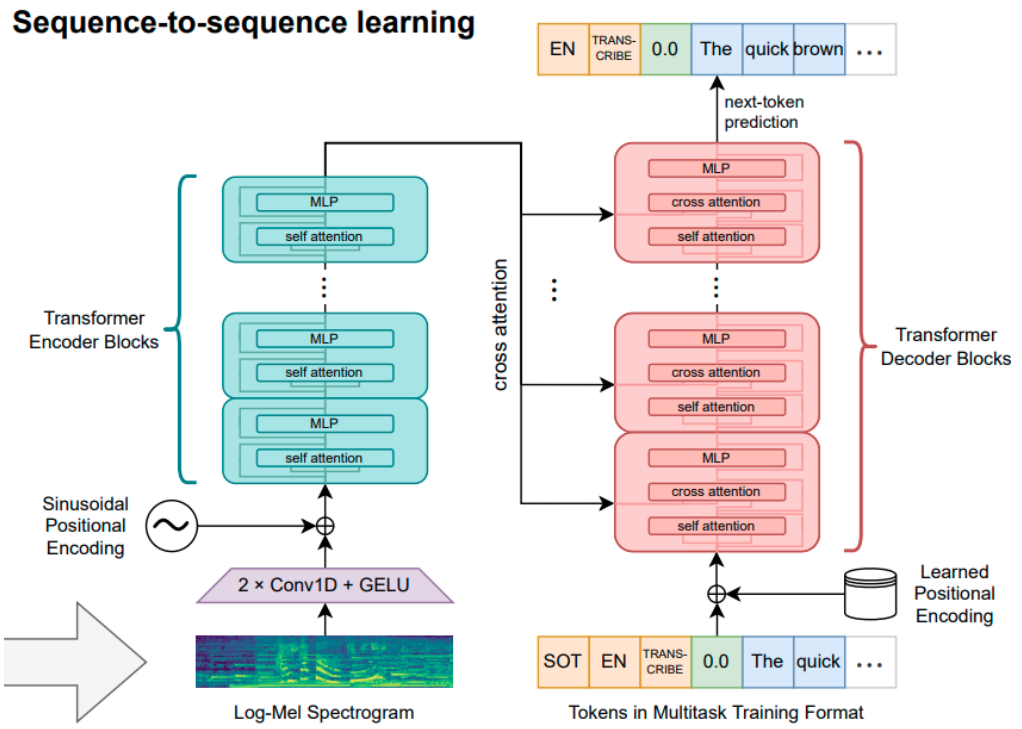

Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
...
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Learned Positional Encoding
SOT
Log-Mel Spectrogram
Tokens in Multitask Training Format

Figure 5.1: Structure of whisper model[45].

Whisper takes audio recordings, divides them into 30-second chunks, and converts the audio signal into Log-mel spectrograms, processing them one by one. Next, the positional encoding of the input is obtained, and the resulting vector serves as the input to the encoder. The model comprises several encoder blocks followed by their corresponding decoder blocks. Various models are available, each varying in size (tiny, base, small, medium, large), chosen to accommodate space and computational constraints. In this study, we have employed the Whisper-tiny (39 M), Whisper-base (74 M), and Whisper-small (244 M) models.

5.1.2. Wav2vec 2.0

Wav2Vec 2.0 is a self-supervised learning framework for speech representation developed by Meta. In this model, four elements stand out from the rest, which are the feature encoder, context network, quantization module, and the contrastive loss.

The first of these model elements that comes into play is the feature encoder, whose goal is to reduce the dimensionality of the input audio data. To achieve this, the feature encoder is composed of a convolutional neural network with 7 blocks, each of which consists of a 1D convolutional layer, a normalization layer, and a GELU activation function[1] [23].

[1] The Gaussian Error Linear Unit (GELU) is a non-linear activation function that amalgamates features from different activation functions, including ReLU, which scales the input value by zero or one; dropout, which stochastically scales the input value by 0; and zoneout, which stochastically scales the input value by one. The central concept is to furnish an activation function that retains non-determinism while preserving dependency on the input value. This results in an activation function that simultaneously maintains non-determinism and input-value dependency. In GELU, the input value is scaled based on how much larger it

In this way, the waveform, after being normalized to a mean of 0 and unit variance, enters the latent Feature Encoder, where it is transformed into a latent feature vector of 512 dimensions. It's worth noting that the audio data is encoded at a 16kHz sample rate.

The next element that comes into play is the quantization module, which converts values from a continuous space into a finite set of values in a discrete space. This element becomes important because speech has a continuous nature. However, when we focus on a language, it has a finite number of phoneme pairs. Hence, the idea of creating a kind of codebook arises, containing all possible pairs of phonemes in a language, which is also a finite number. However, since the number of all possible sounds is enormous, for training convenience, Wav2Vec 2.0 creates G codebooks, each composed of V keywords. So, for quantization, the best keyword from each codebook is selected, and then the vectors from each codebook are concatenated and processed with a linear transformation.

To choose the best key from each codebook, Wav2Vec 2.0 employs the Gumbel-Softmax distribution, a continuous distribution with the property that it can smoothly adapt to a categorical distribution [24]. This distribution offers advantages over the softmax in this context, such as randomization. The model is more inclined to select different code words during training and subsequently update its weights. Additionally, the temperature parameter diminishes the impact of randomization over time. Thus, the final quantized vectors are obtained in this manner.

The next element to consider in the training process of Wav2Vec is the Transformer encoder, which is regarded as the core of Wav2Vec 2.0. It takes the latent feature vectors as input, passes them through a layer called feature projection, which expands the dimensionality of the vectors from 512 to the input dimension expected by the Transformer blocks (768 in the base version of Wav2Vec 2.0 or 1,024 in the LARGE version). Subsequently, it processes this input through 12 Transformer blocks in the base version and 24 in the large version. It's noteworthy that in this interpretation of the Transformer used in Wav2Vec 2.0, positional information is not added to the input vectors, as is done in the original Transformer. Instead, a new grouped convolutional layer is used to autonomously learn relative positional embeddings.

The final element to consider in Wav2Vec is the training process, which consists of two phases. Firstly, there is the pretraining phase, during which the model is trained to learn the latent representation of the input audio. In this phase, a contrastive loss is employed to measure the similarity between the predicted quantized representation of the audio data and the actual quantized representation. Pretraining is carried out using a technique called masking, wherein a random portion of the audio data is removed. The masked audio data is then fed into the model, which is responsible for predicting the missing data. This way, the model must learn to identify the important features of the audio data to predict the missing portions.

Secondly, there is the fine-tuning phase, in which the model is fine-tuned using a Connectionist Temporal Classification (CTC) loss function with a labeled dataset. This loss function minimizes the errors between the predicted transcription and the actual transcription. This phase is employed to enhance the model's performance in the speech recognition task by learning the correspondence between audio data and transcriptions. [7] A diagram of the system can be seen in the figure 5.2.

is in comparison to other inputs. This characteristic introduces smoothness to the activation function when contrasted with ReLU.

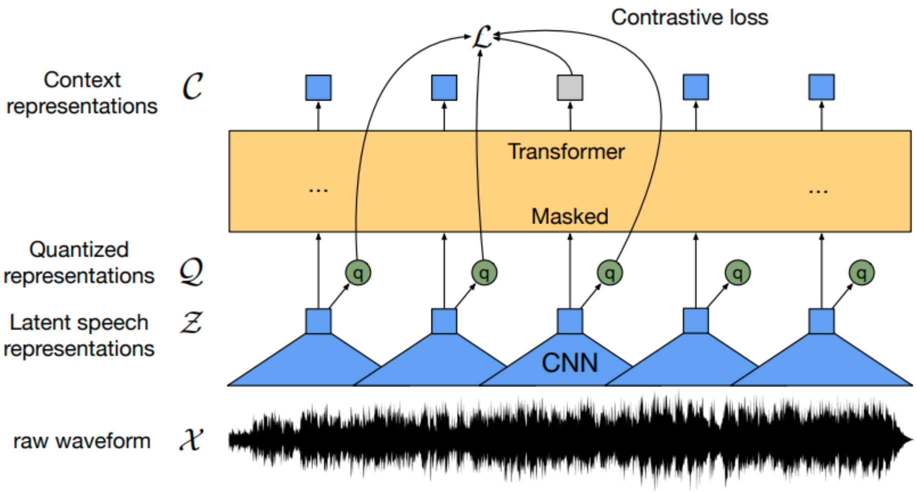


Figure 5.2: Structure of wav2vec2.0 model[7].

5.2 Finetuning

Fine-tuning represents the procedure of adapting a pre-trained model to a new task or domain through continuous training with fresh data. This approach has demonstrated its effectiveness in enhancing model performance.

Within the process of fine-tuning models for speech recognition, the initial critical step involves dataset preparation using the fine-tuning data and aligning it with the model. In this context, specific fields from the available data are carefully selected, which, in the context of this particular study, are limited to audio and its corresponding transcriptions. Subsequently, it becomes imperative to resample the audio to a specific sampling frequency, typically set at 16 kHz, to adhere to the model's specifications.

Furthermore, within the fine-tuning procedures, it is customary to execute the 'Chunking' process, which involves segmenting audio files into smaller units. This partitioning serves to optimize memory usage while concurrently increasing the volume of data available for training. Afterward, is the turn of the tokenization phase, encompassing the segmentation of transcriptions into paragraphs and sentences, further divided into smaller units known as tokens. This streamlined approach facilitates a more effective assignment of meaning to individual text components.

Once transcriptions have been tokenized successfully, the next step is the feature extraction phase, a fundamental step that transforms textual data into a more structured format, thereby facilitating the identification of relevant features. These features hold paramount significance in empowering the machine learning algorithm to enhance its performance.

Finally, after processing the dataset, the retraining of the model starts, for this phase the prepared data is used. The primary objective lies in adjusting the model's weights in such a way that, following this second training phase, its speech recognition capabilities experience an improvement, particularly in the context of speech from individuals with aphasia. This improvement is achieved by leveraging the foundational training of the pre-trained model, which had already demonstrated the ability to transcribe the speech of the majority of the population.

The figure 5.3 shows a diagram of the finetuning process explained above.

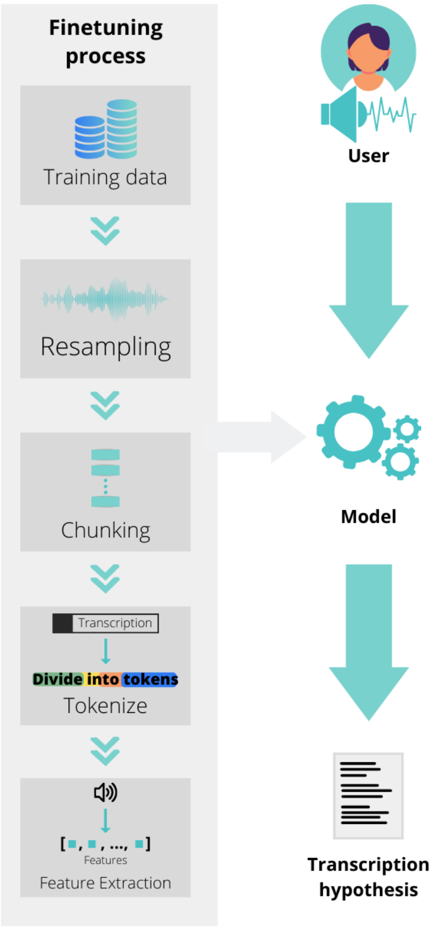


Figure 5.3: Diagram of the Fine-tuning process of an ASR System.

Recently, the term 'foundation models' [10] has been coined to refer to models trained on a large amount of data, which can be adapted to a wide variety of tasks. Due to the significant amount of data and the high computational capacity typically required to train these models, the approach of acquiring a foundation model as a base and then carrying out a specific fine-tuning process for the desired task has become increasingly popular.

In this way, the inherent characteristics of foundation models are leveraged during their pretraining, and their capabilities are adjusted to the specific model during the fine-tuning process. This allows for promising results to be obtained using extensive architectures with a more manageable amount of data.

The use of foundation models has demonstrated promising results in most fields of artificial intelligence. In the case of human-computer interaction and, particularly, in the application aimed at improving communication for people with disabilities, the goals pursued by this project, it is no exception.

CHAPTER 6
Image captioning

Image description generation models are systems that, upon receiving an image, possess the capability to construct a description of its content. In the context of this study, such models emerge as valuable tools to complement the fine-tuning of Automatic Speech Recognition (ASR) systems, potentially making a substantial contribution to a deeper understanding of what users, particularly those afflicted with medical conditions such as aphasia, are endeavouring to communicate or convey. This augmentation in the level of patient comprehension stands out as a noteworthy facet.

It is worth observing that this broadening of objectives entails a subtle shift in direction. The exclusive pursuit of achieving the most accurate transcription of a user's verbal expressions is no longer the sole aim; rather, an additional stride is being undertaken. Within this fresh approach, the goal is to discern the intent behind the verbal utterances of individuals grappling with communication challenges. Moreover, there is a drive to grasp the concepts or ideas they are alluding to. This expanded approach mirrors a quest for a more profound and contextually nuanced understanding of the communication exhibited by individuals constrained by impediments in verbal expression.

6.1 Model discussion

Automated generation of textual descriptions for images stands as a dynamic realm of study that merges computer vision with natural language processing. In its early stages, this domain's primary investigations relied upon associating images with captions grounded in rudimentary attributes [51]. Initial image caption algorithms employed a variety of classifiers to distill features, subsequently employing lexical frameworks or specific templates to formulate descriptive text. Regrettably, this original method exhibited limited adaptability.

Over time, methodologies have evolved, with contemporary approaches embracing deep learning to craft inventive captions. Recent works have delved into diverse deep learning techniques for image captioning. A prevailing strategy involves an encoder-decoder paradigm. The encoder employs a convolutional neural network to distill key image features, converting them into vectors. In contrast, the decoder, typically leveraging Recurrent Neural Networks, generates descriptions through amalgamating image vector features with semantic information. This approach struggled to effectively transmit vital feature information, conveying all details at the initial decoding stage [57].

Moreover, with the emergence of attention models, these have also been applied to the field of image captioning. For instance, Zhou et al. [62] introduced a "text-conditional attention" mechanism, enabling the model to concentrate on specific image attributes based on the hitherto generated text, consequently heightening captioning efficacy. Other attention strategies emerged, such as the "joint attention mechanism," employing multiple LSTMs to concurrently explore diverse image regions derived from visual concept samples, as exhibited in [15]. Another approach involves an attention-based image captioning model, utilizing a pre-trained CNN to abstract image features, coupled with an RNN to formulate captions, as proposed by Agrawal et al. [3].

Furthermore, there is substantial evidence suggesting that the integration of visual information into automatic speech recognition systems can yield a noteworthy enhancement in their precision. This phenomenon is exemplified by the research of Oualil et al. [42], who demonstrated that introducing visual context derived from images led to a 7.8% enhancement in the accuracy of an RNN-based language model's speech recognition capabilities. As previously mentioned, various other studies, including the work explained in [30], have also explored the use of image captioning data to facilitate error correction subsequent to the speech recognition process.

6.2 Image Captioning models

As previously mentioned, the automatic generation of descriptions for images is a complex challenge. It's not merely about recognizing and listing elements within an image, but rather comprehending in some manner what is occurring within the captured scene. This involves considering the spatial relationships among the various elements and being capable of grasping the context presented in the image.

The process of image captioning extends beyond the mere classification of visible objects in the image. It necessitates the ability to describe both the elements present in the image and the interactions unfolding among them. In this regard, the system must not solely identify the objects, but also capture the relationships and dynamics connecting them.

Consequently, within this exploration of models, we won't solely scrutinize approaches exclusively aimed at image captioning. Instead, we'll also delve into multimodal models that encompass image description along with other forms of information.

In the current landscape of image captioning, there exists a diverse range of models, all dedicated to the core objective of describing images. These models not only strive to comprehend visual information but also excel in articulating it coherently and understandably in natural language. Among the notable options is BLIP [35], a framework that employs noisy web data for initializing captions and subsequently filtering out lower-quality instances. BLIP stands out with its cutting-edge performance in tasks like text retrieval from images, generating image captions, and answering visual questions (VQA).

Furthermore, a recent enhancement of BLIP has emerged, known as BLIP2 [34]. This model employs a Lightweight Querying Transformer, which undergoes a two-stage pre-training process. In the initial stage, it focuses on acquiring visual representation, starting from a pre-trained and frozen image encoder. As for the second stage, it commences with a large language model (LLM) that is also frozen, aiming for generative learning of the relationship between vision and language. Consequently, this model manages to achieve outstanding results in the realm of vision and language, despite having significantly fewer trainable parameters compared to similar models.

Additionally, other popular models, such as CLIP-ViT, leverage visual encoders and language models like GPT as decoders to achieve automatic image description, as evidenced in [36].

Furthermore, recently, visual language models have demonstrated promising results in terms of image interpretation. These multimodal systems showcase versatility by incorporating images to respond to queries, generate descriptions, and extract key elements, among other functionalities. A prominent example in this category is MiniGPT-4 [63], a model introduced in 2023. Anchored in a frozen visual encoder and a language model named Vicuna, MiniGPT-4 demonstrates a wide range of capabilities. By aligning these components through a single projection layer, MiniGPT-4 excels in tasks ranging from generating intricate image descriptions to language translation and informative responses.

On the other hand, Flamingo [4] is a Visual Language Model (VLM) created by DeepMind, which can receive an image or video and answer questions based on its information. In addition to visual question-answering, its tasks include scene or event description, multiple-choice visual question-answering, and it also supports few-shot learning and prompting the model with task-specific examples. Flamingo integrates Language Model (LLM) with visual representations - each of them pre-trained and separately frozen - by incorporating intermediate components. This model has been trained on multimodal data. It is noteworthy that the largest version of the model boasts up to 80,000 million parameters.

Another noteworthy model is Otter [31], referred to as "A Multi-Modal Model with In-Context Instruction Tuning." Otter's proficiency is cultivated through training on the novel MIMIC-IT dataset, encompassing a vast collection of multimodal instruction-response pairs. This dataset enhances Otter's competence in tasks involving instruction-following and in-context learning. Notably, Otter showcases exceptional aptitude in tasks that require multimodal perception, reasoning, and contextual comprehension.

In a similar vein, LLaVAR [61] is an enhanced model designed to refine interaction between Large Language Models (LLMs) and humans via instruction tuning. This process fine-tunes a language model's behavior based on specific instructions to align its responses with human expectations. Recent progress integrates images as visual inputs alongside textual instructions, allowing models to respond to image-based prompts. LLaVAR enhances the existing visual instruction tuning pipeline by incorporating text-rich images. By combining Optical Character Recognition (OCR) tools with image captions, LLaVAR prompts a GPT-4-based model to generate conversations involving text-rich images, exemplifying advancements in the synergy of text and images.

CHAPTER 7
Large language models

Once the transcriptions from the fine-tuned model and the patient's contextual description obtained from the image captioning system have been acquired, these will be used as input for a Language Model (LLM) that will provide the final transcription for the system. In this case, the strategy of employing prompting has been chosen to supply the outputs of the two preceding systems, namely, the ASR and the image captioning, to the LLM.

LLMs (Language Models) are pre-trained models with billions of parameters that employ deep learning techniques to process and comprehend natural language. These models are trained on vast amounts of data, enabling them to handle a wide range of tasks based on the provided prompt. They can even perform tasks for which they were not specifically trained by providing several examples of the desired task, a concept commonly referred to as Few-shot learning [19].

7.1 Evolution of LLM

The initial strides in language model research relied on statistical approaches employing Markov models. Nevertheless, they encountered a significant challenge due to the inherent complexity of languages, making it difficult to estimate probabilities for a large number of transitions.

Over time, a fundamental shift occurred in this domain with the adoption of neural language models, which used neural networks to represent sequences of words. Initially, for natural language processing tasks, recurrent neural network (RNN) models were employed. However, it was discovered that these models struggled to capture long-term relationships in text.

As an evolution of recurrent networks, Long Short-Term Memory (LSTM) networks were introduced. The key feature of LSTMs is their ability to retain and recall information over extended periods. This is achieved through the use of memory units known as "memory cells." Each LSTM comprises three essential elements: the "Forget Gate," which decides what to retain or discard in the memory cell, the "Input Gate," which determines what data to incorporate, and the "Output Gate," which dictates which information will be used to generate the network's output at that moment, controlling what is conveyed to the subsequent step in the sequence.

On the other hand, there are Bidirectional LSTM (BLSTM) networks that process the input sequence in two directions: forward and backward.

However, in 2017, a significant breakthrough occurred with the introduction of the Transformer model. This model introduced an attention mechanism that allowed it to capture long-term dependencies by considering the impact of each token in the input sequence in relation to the current token. Furthermore, this structure was highly parallelizable. These features elevated the Transformer model to the state of the art in natural language processing and have propelled most of the recent research in the field, which is based on extensions and enhancements of the Transformer model.

7.2 LLM model discussion

In contemporary discourse, when reference is made to "large language models," it is generally in the context of architectures based on the Transformer framework.

Among the most well-known language models is BERT (Bidirectional Encoder Representations from Transformers) [17], developed by Google in 2018. BERT is trained bidirectionally, enabling it to grasp the context of words in both directions. This has proven highly effective in tasks related to text comprehension and text generation. In the case of BERT, it consists solely of the encoder from the Transformer architecture. One of the most well-known models is GPT-3 (Generative Pre-trained Transformer 3), as reported in [59], released by OpenAI in 2020. which left a profound impression on the community due to its remarkable 175 billion parameters, establishing a new benchmark for the scale of language models.

In 2020, Google also launched its T5 (Text-to-Text Transfer Transformer) model [46]. This versatile model is capable of performing various tasks, including text summarization, language translation, and text classification. Furthermore, T5 was trained on the C4 dataset, a vast corpus comprising 750 GB of clean English text gathered from the internet

Moreover, in 2022, OpenAI introduced ChatGPT, a specialized adaptation specifically designed for dialogue and question-and-answer systems, utilizing the GPT-3.5 model [59]. This particular variant has played a crucial role in increasing awareness regarding the capabilities of such models, both in the academic community and throughout the broader industry.

In 2022, Google introduced the LaMDA [52] dialogue model, which is composed of an astonishing total of 137 billion parameters. This model possesses the capability to generate responses for both text-based and image-based inputs across various contexts and styles.

In that same year, Google launched the PaLM [14] (Pathways Language Model), specializing in natural language processing tasks and featuring an impressive 540 billion parameters. This model holds the potential to serve as a foundation for various use cases in this field.

It is also relevant to mention LLaMA [54], a model trained on texts in 20 different languages, emphasizing the diversity of linguistic applications.

In the year 2023, Meta, in collaboration with Microsoft, unveiled the successor to the natural language processing model known as LLaMA, named LLaMA2 [55]. The largest version of this model boasts up to 70 billion parameters. Furthermore, in comparison to its predecessor, LLaMA2 was trained using a dataset that is 40% larger.

One of the noteworthy improvements in LLaMA2 is its enhanced ability to consider a broader context in natural language processing. The length of context that this new model can comprehend and utilize has doubled in comparison to its predecessor. Additionally, this new model has adopted the 'grouped query attention,' an interpolation of multi-query attention that enhances its speed, as well as multi-head attention, from which it derives its quality.

Another recent model is Falcon [1], an open-source model that offers an option with an impressive 108,000 million parameters. One interesting aspect of Falcon models is their utilization of multi-query attention, which, instead of individual keys and values for each head, shares a single key and value across all heads.

With a similar objective to ChatGPT, Google introduced BARD, a conversational model specialized in responding to questions, which previously relied on LaMDA, but its new update, already available in some countries, has been upgraded to use PaLM2 [6].

Currently, in 2023, models with multimodal capabilities are emerging, allowing the processing of not only text but also a variety of input data types, such as images, audio, and video. Among these models, GPT-4 [40] by OpenAI stands out for its ability to accept inputs of both images and text, generating text-based responses with remarkable precision. It can perform tasks like text generation in different styles, summarization, translation, song composition, and responses to complex questions.

Furthermore, Google has developed PaLM2 [6], an evolution of its original model, enabling tasks such as natural language comprehension, generation, and translation, as well as generating code, audio, video, and images, among others.

CHAPTER 8 Task description and proposed solution

This chapter presents a meticulous description of the task undertaken, providing a deeper understanding of the intricacies involved. The selection of the data to be used, which is a critical aspect, is explained in detail, specifying the criteria and methodology employed in the selection of a representative data set.

In addition, a detailed analysis of the approach adopted to tackle the task is provided.

8.1 Datasets

In order to adapt a previously trained model to the specific characteristics of the distinct group comprised of individuals facing pronunciation difficulties, it is imperative to possess a suitable corpus encompassing an ample collection of speech recordings originating from individuals afflicted by such issues, accompanied by their corresponding transcriptions.

Following a comprehensive investigation, the repository named "talkbank" was successfully identified. In the current context, said repository has been utilized to execute the fine-tuning process, aiming to tailor the model to the unique demands of this situation. Talkbank [37] was originally a project led by Carnegie Mellon University, supported by a collaborative network of hundreds of contributors and numerous collaborators. It encompasses repositories spanning 14 distinct research areas, all of which can be accessed through the links provided on its platform with the objective of advancing fundamental research in the realm of human communication, focusing particularly on spoken language. This repository contains data apported as contributions made by hundreds of dedicated researchers worldwide, getting data in more than 34 languages.

Between the several research areas data included in talkbank in this work Aphasia Bank [38] has been used. Aphasia Bank is a database with recordings and transcriptions in several languages such as Cantonese, Croatian, English, French, German, Greek, Hungarian, Italian, Japanese, Mandarin, Romanian and Spanish conceived for the study of communications in people who suffers from aphasia.

8.2 Data selection and structure

In relation to the data from the Aphasia Bank, both Spanish and English speakers' datasets will be selected. The Spanish dataset comprises four videos, each approximately 40 minutes long, while the English dataset consists of 48 videos, also each lasting around 40 minutes. It is important to note that despite the variation in the number of videos, all of them follow a similar structure.

Each video consists of three main parts. In the first part, the aphasia patient shares their experiences and journey with the disease. The second part involves the participant describing various images, while the third part revolves around the patient discussing the story of Cinderella.

The audio content is recorded in a dialogue format, and the transcription is provided in .cha format. Each sentence is tagged with metadata, including information about the speaker (patient or interviewer), language modifications coded in special tags, timestamps, and POS (part-of-speech) tagging. Additionally, for the Spanish dataset, a word-by-word English translation of each sentence is included.

During the initial phase of data processing, we will focus solely on the sentence transcriptions, disregarding other metadata. Furthermore, we will specifically use the sentences spoken exclusively by the patients with aphasia to train the proposed system, excluding the utterances voiced by the interviewer. As a result, the total audio length obtained from each video will be halved, resulting in approximately 15-20 minutes per video.

To achieve this, we will employ several scripts with regular expressions to filter and adapt the transcription format, ensuring it aligns with the desired format suitable for training the selected models. Furthermore, the complete recording will be divided based on the transcribed sentences and their respective segments of audio files. Subsequent to this segmentation, the pathway to the sub-audiofiles along with their corresponding transcriptions will be preserved within a csv file. This meticulous process ensures that the model receives effective training with relevant patient speech data, enabling it to effectively address the specific challenges associated with aphasia.

By using only the pertinent sentences and optimising the audio duration, we created a focused and tailored dataset that will enhance the model's ability to understand and interpret speech patterns unique to individuals with aphasia.

8.3 Approach

The proposed approach involves the fusion of different models to address the task.

In the context of this multimodal proposal, three essential components are considered: a speech recognition model, an image description model, and a large language model. The diagram in Figure 8.1 provides a visual representation of the entire process.

Firstly, audio recordings are acquired, which are adapted to the input format of the model. These recordings will be used as a dataset for fine-tuning the selected speech recognition model.

On the other hand, there are also images representing the user's visual context. These images are used as input for an image captioning system. Here, two different approaches are explored: in one, the image captioning system generates a textual description of the image, while in the other, it produces a set of keywords. Subsequently, the performance of both approaches is compared.

Finally, both the audio transcription obtained from the speech recognition system (after adjusting the weights during fine-tuning to adapt to the set of individuals with aphasia) and the description or keywords (depending on the approach) of the user's contextual image are input into a large language model. This model determines whether adjustments to the transcription are necessary using the visual context information obtained from the image description.

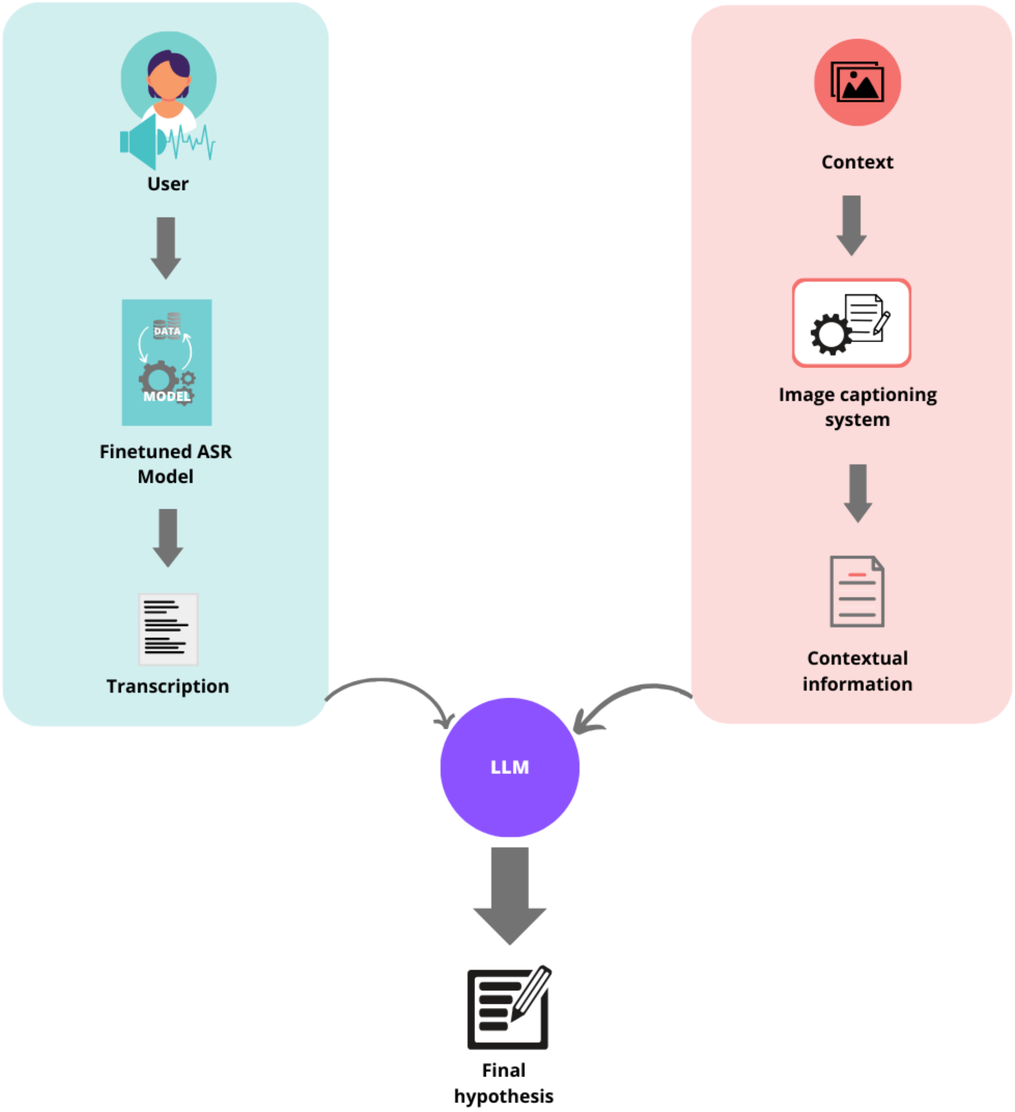


Figure 8.1: Diagram of the complete system.

Upon completing the development of the solution, a comprehensive analysis of the obtained results will be conducted. This analysis will be divided into two distinct approaches to assess the effectiveness of the solution. Firstly, an evaluation will be carried out using solely the speech recognition model (ASR) that has undergone fine-tuning. During this phase, a comparison among different ASR models will be conducted, and experimentation with various hyperparameters will aim to achieve the best possible results in the fine-tuning process.

On the other hand, the previously described complete system will be assessed, which incorporates the interpretation of environmental descriptions generated by the image description system. This comprehensive system will be responsible for refining the transcriptions suggested by the speech recognition model based on the environmental descriptions derived from the images.

CHAPTER 9
Evaluation and results

In this section of evaluation, we conduct a series of experiments wherein we adapt various systems of different sizes to the specific domain of models dealing with aphasia. Our objective is to ascertain whether this model adaptation confers an advantage and leads to improvement compared to the baseline models. Our focus remains on the specific subset of individuals with aphasia.

To implement these experiments, we employ the programming language Python and leverage the open-source Tensorflow library, developed by Google, which enjoys substantial recognition in the field of machine learning.

Our process commences with the retrieval of data from Aphasia Bank. As elaborated in section 8.2, we procure 4 videos available in Spanish and 48 videos available in English. These videos, accompanied by transcriptions in .cha format, serve as our primary sources of information.

Following the data retrieval, we create and execute a Python script to refine the transcriptions, extracting only the relevant information from the .cha files. Additionally, we convert the videos into audio files in a suitable format in accordance with the model's requirements, whether it be wav or mp3. We segment the videos based on time markers corresponding to the phrases, ensuring precise alignment between each audio snippet and its associated transcription. This processed information is stored in a CSV document containing two columns: one indicating the path to the specific audio file within the system and the other containing the corresponding transcription text. Given that we are dealing with interviews, we choose to exclude the audio of the interviewer, concentrating exclusively on segments related to individuals with aphasia.

Subsequently, we randomly shuffle the rows of this CSV and partition it into three distinct dataframes: "train" containing 75% of the data, "test" with 15%, and "validation" with 10%.

These data undergo transformation into Datasets and undergo a process of tokenization, processing, and feature extraction. In order to do this step, the tokenizer, processor and feature extractor of the corresponding model and language in huggingface are used for the transcription task. Finally, we utilize these datasets in the training process to adapt the pre-trained model to the specific characteristics of the group of individuals with aphasia.

To evaluate the effectiveness of our adapted models, we implement another script to compute the Word Error Rate (WER) based on transcribed sentences from the validation set and the original transcriptions of the sentences.

The initial data partitioning is approached from two different perspectives. Firstly, we consider a context-independent approach where data segmentation is based on the thematic or contextual content of the individuals' speech. This segmentation allows us to examine the impact of fine-tuning on transcriptions across various topics. In contrast, we also implement a speaker-independent approach by dividing the data based solely on the speaker's identity, disregarding the underlying context or subject matter of the discourse.

Additionally, we delve into the influence of context information on transcription quality. This entails evaluating the potential enhancement in transcription accuracy when context is incorporated into the analysis. Across this chapter, we will present and analyze the outcomes of the various test-case experiments.

9.1 Metrics

The Word Error Rate (WER) metric is commonly used to measure the speech recognition models' performance. WER is calculated as follows:

$$WER = \frac{(Substitutions + Insertions + Deletions)}{RefWords} \tag{9.1}$$

The WER metric involves comparing the model's hypothesis with the original transcription and determining the number of substitutions, insertions, and deletions required to match the reference sentence. This value is then divided by the number of words in the reference sentence. Substitution occurs when a particular word is misconstrued, leading to the use of an alternative word in its place. Insertion, on the other hand, involves the addition of a word or phrase that was not originally spoken. An example of insertion is when a single spoken word is mistakenly transcribed as two separate words. Lastly, deletion takes place when a spoken word is entirely omitted from the transcription. A lower WER indicates a better-performing model.

9.2 Context independent Tests

Within the realm of context-independent testing, the procedure involves extracting part of the second of the three segments into which the videos are divided. In this specific segment, participants undertake the task of describing multiple images, with a particular focus on the segment where two specific images are being described. This entails the extraction of this video segment along with its corresponding transcription from all the videos that constitute both the training and testing sets. This methodology ensures that the model validation and metrics are computed based on a dataset and context that have not been previously encountered. It is noteworthy that this step is undertaken initially, even prior to the transcription cleaning or the creation of the CSV file containing the videos transformed into audio and segmented.

Subsequent to this segmentation, the process mentiones at the begining of this chapter is executed to prepare the CSV file and the distinct datasets (training and testing). These datasets will be employed for the fine-tuning of the pre-trained model. Furthermore, the same procedure will be applied to the audio files and corresponding transcriptions of the description of these isolated images, which together constitute the validation set.

It is worth emphasizing that the decision to reserve the portion of dialogue where images are described for validation was not arbitrary. Given that we possess the pertinent images being described, they can be employed as context for patients with aphasia in the second phase of this endeavor. In this subsequent stage, we will harness models designed for generating descriptions of images, enabling us to articulate what the patient is perceiving. In this context, we are specifically referring to the images discussed within the validation set. By integrating this visual context with the transcriptions generated by the pre-trained model, we can subsequently incorporate this contextual information into a language model. Through this approach, we can delve into the extent to which environmental information contributes to the comprehension and transcription of what individuals with aphasia are attempting to convey.

Once the data has been suitably prepared, partitioned, and tokenized, and a training and a test dataset have been created, the pivotal stage emerges: defining the training parameters for the fine-tuning process of a pre-trained model using the Hugging Face Transformers library. In this phase, an array of hyperparameters will be established to steer the model optimization and its tailoring to the specific task.

Among these parameters, the output directory is included, serving as a repository for trained models and other pertinent elements. Furthermore, the batch size for training is determined, influencing the quantity of examples processed in each iteration. A pivotal hyperparameter is the learning rate, which governs the magnitude of weight adjustments within the model in each training cycle. Essentially, the learning rate designates what fraction of the gradient is used to update the model's parameters at each step.

Additionally, the warmup steps are set, governing the gradual increase of the learning rate from zero to its defined value. The max steps denote the maximum number of training iterations and are fundamental in controlling the duration and scope of the process.

In parallel, the frequency at which steps are saved and evaluated during training is established. This is indispensable for the constant monitoring of progress and the selection of the best model. Moreover, the maximum length of generated sequences during the evaluation phase is defined – a pivotal factor in controlling coherence and the extent of predictions.

With these hyperparameters defined, an object is constructed to encapsulate both these values and related configurations. This object is enriched with the inclusion of the pre-trained model, training and test datasets, the data collector, the function responsible for calculating pertinent metrics, and the appropriate tokenizer for the given task.

Finally, when all is in place, the training process is triggered by invoking the "train()" method of the trainer object. Throughout this process, the model adapts to the training data, its weights are fine-tuned through backpropagation, and its performance is evaluated at specific intervals, thus providing continuous feedback on the quality of predictions.

Comprehensive experiments were conducted to ascertain the optimal parameter values for our model. Specifically, we investigated the effects of different values on the model's performance in terms of WER. One of the critical parameters under scrutiny was the batch size employed during training.

Due to constraints related to data size, model capacity, and GPU size, it was determined that the maximum viable batch size was 8, as this value allowed computations to fit within the GPU. Moreover, through various trials, it was observed that a larger batch size tended to yield improved outcomes in our case. Therefore, considering the GPU-imposed limitation, we opted to maintain a consistent batch size of 8 for all subsequent experiments.

It's worth noting that, owing to time restrictions, hyperparameter tuning was initially executed on the "whisper-small" model. Subsequently, these same hyperparameters were applied to the training configurations of other sizes and models. The detailed results for different hyperparameter values are presented in Table 9.1.

learning rate	Max steps	warmup steps	WER(%)
10^{-5}	1000	500	72.80
10^{-5}	1500	500	48.27
10^{-5}	2000	500	40.81
10^{-5}	3000	500	33.55
10^{-5}	4000	500	32.43
10^{-5}	1000	300	70.36
10^{-5}	1500	300	36.76
10^{-5}	2000	300	44.48
10^{-5}	3000	300	32.45
10^{-5}	**4000**	**300**	**31.53**
10^{-5}	1000	150	51.05
10^{-5}	1500	150	43.75
10^{-5}	2000	150	45.18
10^{-5}	3000	150	33.09
10^{-5}	4000	150	33.57
10^{-4}	1000	500	152.95
10^{-4}	1500	500	79.76
10^{-4}	2000	500	58.20
10^{-4}	3000	500	59.18
10^{-4}	4000	500	39.27
10^{-4}	1000	300	166.54
10^{-4}	1500	300	54.63
10^{-4}	2000	300	101.24
10^{-4}	3000	300	62.19
10^{-4}	4000	300	38.78

Table 9.1: Adaptation of hyperparameters in the Whisper-Small model using the Aphasia English dataset and a context-independent approach.

In order to analyse the evolution of the parameters and trends in the results of the various tests conducted with different hyperparameters more effectively, we have created a graph that allows us to compare the results visually. As can be observed in the figure 9.1

As evident from both Table 9.1 and Figure 9.1, the most favorable outcomes have been achieved with a training duration of 4000 steps, a learning rate of 10^{-5}, and 300 'warm-up' steps. It is also noteworthy that, although the 'warm-up' step count does not significantly impact the final WER value at step 4000, the choice of learning rate makes a substantial difference, with a more favorable WER attained at the rate of 10^{-5}.

Furthermore, it is interesting to observe that when employing a learning rate of 10^{-5}, the evolution of WER from step 3000 onwards appears to stabilize, showing a more modest improvement. Consequently, this has led to the decision to adopt, for subsequent trials, a learning rate of 10^{-5}, 300 'warm-up' steps, and a maximum of 3000 steps.

Once the most effective values were identified to optimize the Whisper-small model, these values were employed during the finetuning process in other models.

In the experimentation process concerning the fine-tuning of various systems, we have selected three sizes of Whisper models (small, base, and tiny), in addition to the wav2vect2.0 model. For each of these models, we have executed the fine-tuning process previously elucidated in this same chapter, utilizing the data acquired from AphasiaBank. Subsequently, upon obtaining models with adjusted weights, we proceeded to compare the WER obtained using the validation dataset. This metric has been calculated

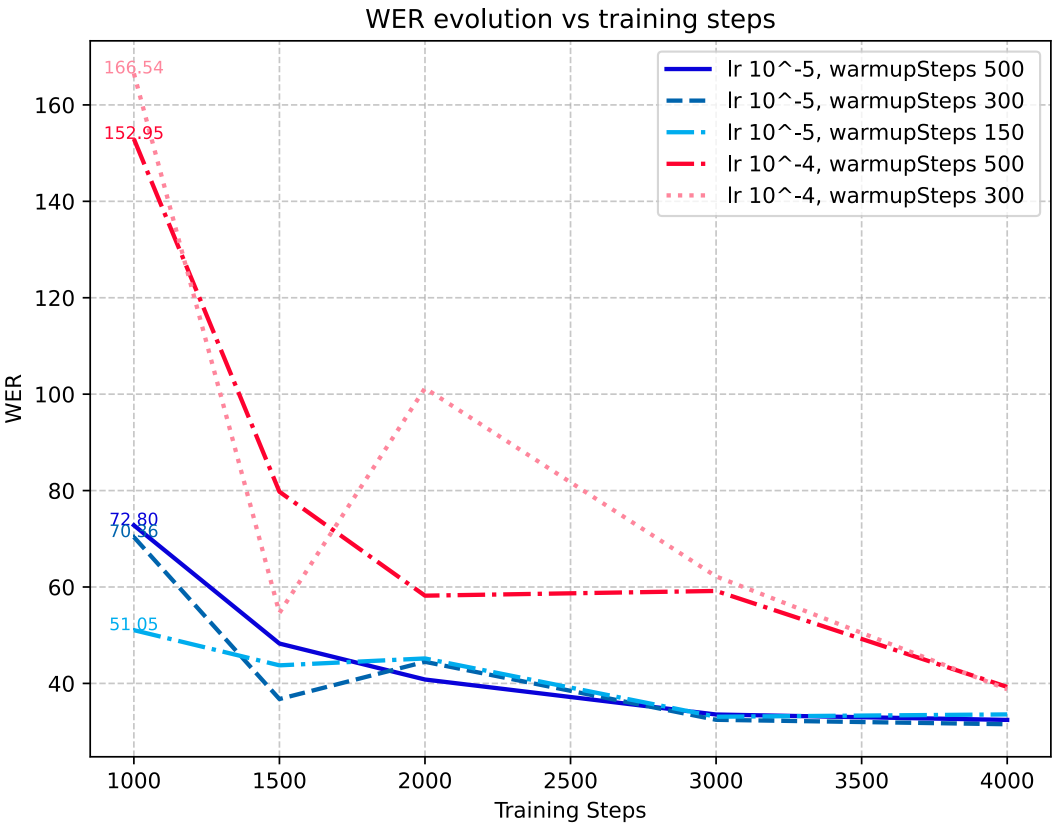


Figure 9.1: WER evolution depending on the selected hyperparameters during the fine-tuning process of the Whisper-Small model.

for both the base models (models without the fine-tuning process) and the models fine-tuned through the fine-tuning process.

The results of WER obtained are presented in Table 9.2. It is imperative to note that these results stem from data partitioned following the content-independent approach.

It is essential to note that the validation set used to calculate the WER metric consists of dialogues different from those encountered during the training phase. Therefore, the validation data represents a distinct audio theme from the training data.

Model	**WER(%)**
Whisper-small	70.36
finetuned Whisper-small	31.53
Whisper-base	52.27
finetuned Whisper-base	31.61
Whisper-tiny	54.69
finetuned Whisper-tiny	45.04
wav2vec 2.0	68.02
finetuned wav2vec 2.0	49.22

Table 9.2: Achieved WER in the experiments with the English test set using the context-independent approach.

The table demonstrates the impact of fine-tuning the models with domain-specific data, as opposed to relying solely on the initial model. Lower WER values indicate improved accuracy in transcribing speech-to-text for the given domain.

As shown in Table 9.2, the Whisper-small model achieved a significant improvement, with the Word Error Rate decreasing from 70.36 to 31.53. This represents a relative WER improvement of 55.23%. On the other hand, for the Whisper-base model, the improvement was from 52.27 to 31.61, resulting in a relative improvement of 39.5%.

It's worth noting that prior to training the various models with the English dataset, comprised of 48 speakers, from whom approximately 20 minutes of recording each will be utilized, tests were conducted in Spanish with just 4 speakers and approximately one hour of audio recordings. In these tests, a direct comparison was made between the pre-trained "small" Whisper model and the pre-trained ESPnet2 model, along with the fine-tuned Whisper model.

The hyperparameters employed for conducting the fine-tuning process of the Whisper model on the Spanish language dataset encompassed a learning rate value of 10^{-5}, a batch size of 8, an initial warmup steps count of 500, coupled with a total of 3000 training steps. The outcomes pertaining to WER for both the pre-trained models and the adapted model are delineated within table 9.3.

Model	**WER(%)**
ESPnet2	182.33
whisper-small	121.57
Finetuned whisper-small	36.34

Table 9.3: Achieved WER in the experiments conducted with the Spanish test set.

As evidenced in Table 9.3, it is apparent that in the case of tests conducted in Spanish, there is also a significant disparity between context-adapted models for individuals with aphasia and merely pre-trained models. For instance, in the case of Whisper, the WER decreased from 121.57% to 36.34%, signifying a relative improvement of 70.13%.

It is imperative to underscore that, in context-independent testing, the WER value may be subject to the influence of specific portions of image descriptions selected from the training dataset used as the validation set. This is because, depending on whether the chosen image includes elements from a more or less specific vocabulary, the metric's results could undergo substantial variations.

As demonstrated in various experiments, particularly in the case of individuals with speech difficulties, the fine-tuning process assumes significant importance. Upon analyzing the test case, it is presumed that the improvement primarily pertains to the end-to-end model component that corresponds to the role played by the acoustic model in hybrid systems. This is because the fine-tuning process facilitates a more precise understanding of the user's speech patterns.

As a result, a second experiment will be conducted in which the division of training, testing, and validation sets will be based on speakers rather than content. This approach aims to further enhance the model's performance in interpreting diverse speech patterns, ensuring greater accuracy and ease of use for users facing speech difficulties.

9.3 Speaker independent Tests

After the initial model training utilizing data splitting based on phrase content encountered certain challenges, an alternative methodology was adopted. In this instance, the training, testing, and validation sets were partitioned according to the individual speaker, ensuring that the sentences in the training set were uttered by a distinct speaker from those present in the validation set. Additionally, efforts were made to maintain speaker balance across the various sets, aiming to include an equal representation of both male and female speakers in the training set as well as in the test and validation one.

In order to obtain the metrics the same validation set has been used for all the experiments. In addition, for this set of tests, more data has been obtained than in the previous tests due to the addition of recordings of other different speakers available in the aphasiaBank. Thus, in this case the training set has around 720 minutes of void while validaition one has around 105 minutes. The training process has been very similar to the one done in the previous section. As for the hyperparameters, their selection has also been based on the criteria outlined in the previous section, keeping the learning rate at 10^{-5}, 300 warm-up steps, and a maximum of 3000 steps.

Model	WER(%)
whisper-small	61.25
finetuned whisper-small	35.60
Whisper-base	176.43
finetuned Whisper-base	133.50
Whisper-tiny	219.73
finetuned Whisper-tiny	43.71
wav2vec 2.0	85.49
finetuned wav2vec 2.0	51.27

Table 9.4: Achieved WER of the experiments with the english validation set in speaker-independent approach.

As depicted in the table, a clear trend is observed in the base models where larger models, characterized by a higher number of parameters, consistently yield better WER metrics in this context. It is noteworthy that, in all cases, the fine-tuned model consistently outperforms its counterparts in terms of WER. We observe that the best WER in this set of tests is also achieved by Whisper-small, with a WER value of 35.60%, achieving a relative improvement of 41.82% compared to the non-fine-tuned Whisper-small model, which obtained a WER of 61.25%.

Furthermore, the substantial improvement in the WER of the Whisper-tiny model is particularly noteworthy, transitioning from a WER of 219.73% to 43.71%. This represents a relative improvement of 80.02%. It is important to highlight that these values were obtained through partitions following a different approach from that used in the previous section. Therefore, the validation set (as well as the training and test sets) is not the same as that used in the previous section, which explains the differences in the results.

9.4 Multimodal System Test

In this section, a series of tests will be conducted to assess the relevance of contextual information in our specific case for enhancing transcriptions. To accomplish this, various tests will be executed using Visual Language Models (VLMs) with a specific focus on extracting information about the image being observed by the user. This procedure will yield contextual information. Conversely, we will utilize the fine-tuned model that has produced the most favorable results. In this particular scenario, given the nature of the test, we will employ content-independent models.

The rationale for this selection lies in the fact that the forthcoming tests will incorporate contextual data sourced from images described by patients in specific recordings contained within the Aphasia Bank. In order to validate the system, it is imperative that the context associated with the aforementioned images has not been previously encountered by the system. This precautionary measure is instituted to mitigate the risk of overfitting.

The effectiveness of this approach is augmented through a meticulously conducted data partitioning process during the training of context-independent models. To elaborate, segments of video content wherein users described the particular images under consideration have been excised from all video recordings. Consequently, these segments were not included in either the training or test sets utilized during the fine-tuning process of the ASR model.

This partitioning strategy entails the extraction of segments from both the training and test datasets, encompassing instances where users provide descriptions of specific images. By adhering to this methodology, the pre-trained system will not have encountered the precise contextual information pertaining to the image designated for validation.

The forthcoming procedure shall entail the following steps: Initially, the pertinent audio files corresponding to a given image will be procured and subjected to transcription. This transcription will be conducted employing the fine-tuned model that has demonstrated superior performance in context-independent experiments. Concurrently, from the available video archive within the Aphasia Bank dataset, the image referenced within the sentences of our validation set will be meticulously extracted. Subsequently, this extracted image shall be introduced as input to our Visual Language Model (VLM). It is of paramount importance to bear in mind that the video material, from which we extract these images, has not been incorporated into either the training or test sets utilized during the model fine-tuning process.

During this phase of employing the Visual Language Models, two discrete approaches will be undertaken. Initially, the VLM will be instructed to generate an inventory of elements depicted within the image. Conversely, in another test, the VLM will be assigned the responsibility of furnishing a detailed narrative elucidation of the image.

Considering that the image under scrutiny follows an illustrated story strip format and encompasses multiple panels, each panel will be presented individually to the VLM. Subsequently, the resultant transcriptions from each panel will be merged.

Upon acquisition of the VLM's response, which may encompass either the list of elements or the image description, this description shall function as a stimulus for a Large Language Model (LLM), in conjunction with the transcription produced by the fine-tuned ASR system. Subsequently, the LLM will be tasked with the identification and correction of errors it discerns with confidence within the transcription.

As previously mentioned, the objective of this experiment differs from previous goals. Instead of focusing on achieving the most accurate transcription, the emphasis shifts towards understanding the intended meaning when individuals with aphasia express a specific phrase. This is because aphasia can lead to various challenges, such as difficulty in finding the right words to convey what patients want to communicate, speech limitations, the use of short phrases, or the lack of coherence in expressions, as explained in section 2.2. In some cases, the phrases spoken by individuals may not accurately convey their intention. Therefore, challenges arise in this context.

The metric that was previously used to compare different transcription systems, the Word Error Rate, is no longer suitable for our current purpose. This is because obtaining a transcription that closely resembles the reference phrase does not guarantee grammatical correctness or a complete reflection of the user's intention. Consequently, to assess the effectiveness of this hybrid system, we have adopted a human evaluation methodology.

Regarding the language model choice, for the sake of convenience and accessibility, two widely recognized models, GPT-3.5 and LaMDA, have been tested.

In the first experiment, a list of keywords was supplied to the process, while in the second experiment, an image description was employed. In both cases, visual information was obtained through the LLaVAR VLM. In addition to the descriptions, the language model was provided with the audio transcription, obtained through the previously fine-tuned ASR system. Specifically, the Whisper-Small fine-tuned system was utilised, as explained earlier, following the context-independent approach. This choice was made because Whisper-Small fine-tuned had yielded the most favourable results in previous experiments. These tests were repeated using GPT-3.5, as well as LaMDA. In Figure 9.2, it is possible to observe a specific example of the process followed by this multimodal approach, for one of the phrases that has been tested in various experiments.

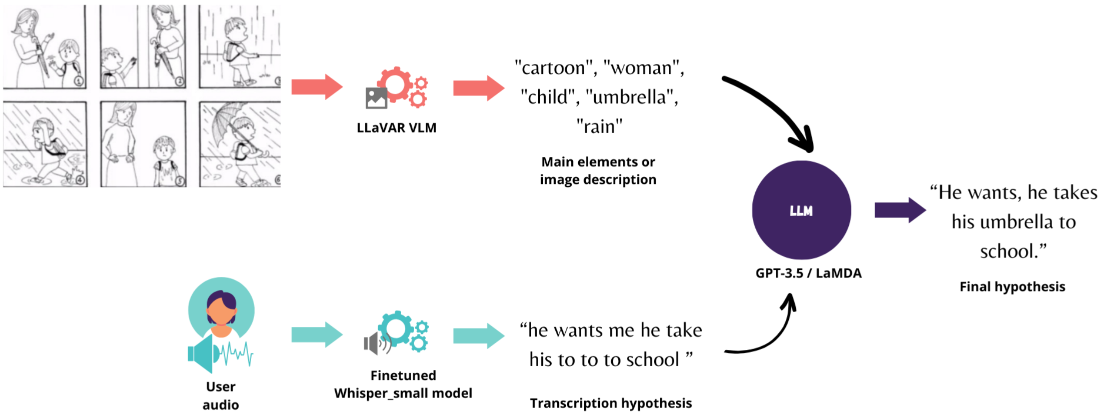

LLaVAR VLM
"cartoon", "woman", "child", "umbrella", "rain"
Main elements or image description
LLM
GPT-3.5 / LaMDA
“He wants, he takes his umbrella to school.”
Final hypothesis
User audio
Finetuned Whisper_small model
“he wants me he take his to to to school ”
Transcription hypothesis

Figure 9.2: Example of usage of the complete multimodal system given a specific image and audio.

To conduct the human evaluation, a survey was administered to various individuals. Each survey question included the audio spoken by a person with aphasia, along with transcriptions generated by the different systems under comparison. These systems include the fine-tuned Whisper model, the fine-tuned model augmented with the list of elements from using GPT-3.5 as LLM, the fine-tuned model augmented with the image description from using GPT-3.5, the fine-tuned model augmented with the list of elements from using LaMDA as the LLM, and the fine-tuned model augmented with the image description from using LaMDA.

Regarding the survey conducted as part of the human evaluation process recordings of audio containing phrases spoken by individuals with aphasia were employed, alongside transcriptions generated by four hybrid systems and a fifth transcription obtained through the ASR fine-tuned system. The survey comprised two questions for each of the seven presented audio segments. The first question aimed to determine which of the provided transcriptions was more similar, in terms of words, to what the speaker expressed in the audio. Conversely, the second question sought to identify which of the transcriptions better aligned with what the speaker genuinely intended to communicate. An example of these questions can be found in the appendix A.

Fifteen different users participated in this survey, providing their responses to the questions posed in relation to the seven audio segments. As for the evaluation of responses, the users' feedback was consolidated for each audio segment, resulting in a model percentage. To achieve this, the number of users who determined that a particular model performed best in each specific question was summed, separately for each model. Subsequently, the value obtained for each model was divided by the total number of votes across all questions to obtain the percentage Table 9.5, shows the survey results based on the first question posed for each audio, where the most similar transcription to the audio was requested.

Model	Result
original transcription	34.25%
finetuned whisper-small	47.95%
LaMDA with keywords	0%
LaMDA with descripcion	0%
GPT 3.5 with keywords	6.16%
GPT 3.5 with description	11.64%

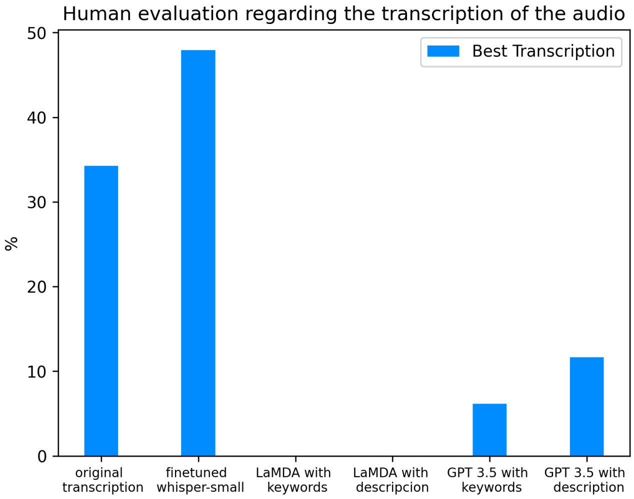


Table 9.5: Results of human evaluation regarding the system that best reflects word-for-word transcription of the audio.

As can be observed in the table 9.5, in the survey, we compared the transcriptions of all the systems mentioned above, as well as the original transcription. It is worth noting that the system rated highest by the surveyed users in terms of the best transcription is the Whisper fine-tuning, with 47.95% of users. It is noteworthy and has surprised us that this system has, overall, received an even higher rating than the original transcription, which was voted as the best system by 34.25% of users.

Regarding the rest of the systems, we observe that, in this case, the systems using GPT as their language model receive higher ratings from users compared to the system using LaMDA. Additionally, for the surveyed users, systems that have incorporated image descriptions seem to provide transcriptions that are closer to their expectations compared to those relying solely on a list of elements present in the image, which we refer to as 'keywords' in the table.

With regard to the second question, in which users are asked to express their opinion about what the speaker is referring to in the sentences, the results vary significantly compared to the previous experiment, even though this question involves the same audio recordings and presents the same options. These results can be seen in the Table 9.6. Furthermore, in the same figure, one can observe a bar chart that visually presents these results.

In relation to this survey question, we have added the same systems evaluated in the previous experiment as possible choices. The inclusion of the original transcription of the audio included in the dataset has been of particular interest in this experiment, with the aim of assessing to what extent a multimodal system can enhance its ability to accurately reflect the possible intention of the user conveyed in the reference phrase extracted from the dataset.

In this context, we observe that the system that achieves the best performance, according to 33.85% of the surveyed users, is GPT 3.5 as a language model, to which audio transcriptions and a list of elements present in the visual environment of the user are added as input. The second most highly rated system turns out to be the one employing LaMDA as a language model, to which transcriptions and a list of the main elements in the image are provided, garnering 19.23% support. The third position is held by the system utilizing GPT 3.5 as a language model, which receives visual information from the environment in the form of an image description, receiving a 13.85% preference from users.

It is noteworthy that, overall, the surveyed users consider that the transcriptions provided by these three models convey the speaker's intention more accurately compared to the original audio transcription found in the dataset. Only 12.3% of the users believe that the original audio transcription is the model that best interprets what the user wishes to express.

Model	Result
Original transcription	12.30%
finetuned whisper-small	10%
LaMDA with keywords	19.23%
LaMDA with descripcion	10.77%
GPT 3.5 with keywords	33.85%
GPT 3.5 with description	13.85%

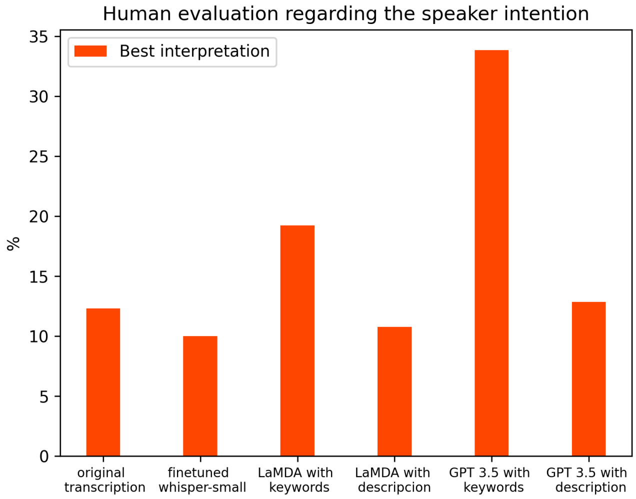


Table 9.6: Results of human evaluation concerning the system that best reflects what the speaker intends to refer to.

As can be observed in this second conducted experiment, generally speaking, the systems to which visual context information has been provided through a list of key elements from the context are the ones yielding the most favorable results. In this case, it is not only the systems using 'keywords' that outperform their counterparts, namely, the systems employing the same language model but with a description, but also the system to which a list of keywords is provided are the ones that users have rated more positively overall.

This phenomenon can be attributed, among other reasons, to the fact that having only the main elements can provide the crucial context information without introducing grammatical information that may confuse the system and alter the initial transcription.

CHAPTER 10
Conclusions

In this chapter, we will summarise the main contributions of this work, the experiments conducted, and the results obtained. Additionally, we will address future work.

The objective of this project has been to provide solutions to enhance user interaction with pronunciation problems, specifically aphasia, in relation to automatic speech recognition systems. To achieve this, experiments were conducted using fine-tuning of various models. Furthermore, a multimodal system was introduced and tested, wherein transcriptions from the fine-tuned models and descriptions of an image representing the user's visual context are used as input to a language model. This model will attempt to generate a sentence that best reflects what the speaker with aphasia tried to convey when pronouncing the original sentence.

To address this project, we began by familiarising ourselves with some of the most common speech disorders, focusing in particular on the case of aphasia, a disorder caused by damage to the brain areas responsible for language, which affects the communication of those who suffer from it, as explained in Chapter 2.

Subsequently, we undertook a thorough search of state-of-the-art projects with objectives akin to ours. Moreover, we carried out a review of the evolution and present-day methodologies employed in the realm of automatic speech recognition. We also introduced the domains of image captioning and large language models, offering an overview of some of the most prominent contemporary models in each domain. Furthermore, we detailed the dataset used in our study.

Regarding the experiments and systems developed in this project, in the first part related to fine-tuning ASR models to adapt them to the subfield of transcribing individuals with aphasia, we pursued two approaches based on data partitioning. On one hand, we performed partitions based on context, and on the other hand, we conducted partitions based on the speaker.

In the experiments carried out in the context-based data partitioning approach, we compared the Whisper-small, Whisper-base, and Whisper-tiny models with Wav2Vec 2.0. Additionally, we fine-tuned these models to tailor them to our specific domain. In the case of the best-performing model, Whisper-small, we achieved a Word Error Rate of 31.53%, representing a relative improvement of 55.23% compared to the base model without the fine-tuning process. In the Spanish version, Whisper-small also experienced a significant relative improvement after fine-tuning, with a reduction in WER from 121.57% to 36.34%.

Finally, in the case of fine-tuning ASR models following the "speaker-independent" approach, the model that yielded the best results among those analyzed was also Whisper-small, with a relative improvement of 42.82% in its training, reducing the Word Error

Rate (WER) from 61.25% to 35.60%. As mentioned earlier, this latter result was the best achieved in this set of experiments.

These results confirm the effectiveness of the fine-tuning process in adapting pre-trained models, which, despite having very low WER rates in the general population, exhibited much more limited performance when used by individuals with speech problems such as aphasia. As we have seen, adapting the weights of these models with domain-specific data from people with aphasia has led to a significant improvement in the error rate, yielding very promising results in this regard.

Furthermore, the work also addressed the objective of not only transcribing but also effectively conveying what users with oral communication problems wish to communicate. Our experiments combined the transcription obtained from the best fine-tuned ASR system with visual information provided by a computer vision model (VLM) and a large language model (LLM).

In these experiments, LLaVAR was used as the Visual Language Model (VLM), Whisper-small fine-tuned as the ASR model, and the results were compared between GPT 3.5 and LaMDA as Large Language Models (LLM). Furthermore, tests were conducted both using the VLM to obtain a transcription of the image as visual context information and using a list of the key elements of the image as visual context, also obtained with the VLM. In this case, human evaluation was employed, and 33.85% of the users specified that, concerning the system that best reflects the speaker's intention, GPT 3.5 using a list of the most important elements as contextual information was the preferred choice. On the other hand, users who participated in the human evaluation were also asked which system they believe transcribes the user's speech word-for-word better. In this case, the most part of the users agreed that the fine-tuned Whisper-small performed best.

In conclusion, the initial objectives of this work have been successfully achieved. We have significantly enhanced the performance of pre-trained ASR systems in their interaction with individuals facing aphasia. Furthermore, we have developed systems that aim to more accurately reflect what users wish to communicate, thereby improving their expressive abilities. The results of human evaluations support the effectiveness of these systems.

The outcomes of this study provide grounds for optimism regarding the potential for future systems to more effectively harness the multimodal information available from users. This could contribute to enhancing communication and fostering the inclusion of individuals with disabilities such as aphasia. Moreover, these results inspire us to continue exploring the use of foundational models, fine-tuning, and the amalgamation of different models to leverage user information and context, with the aim of enhancing the quality of interaction among individuals and with technology.

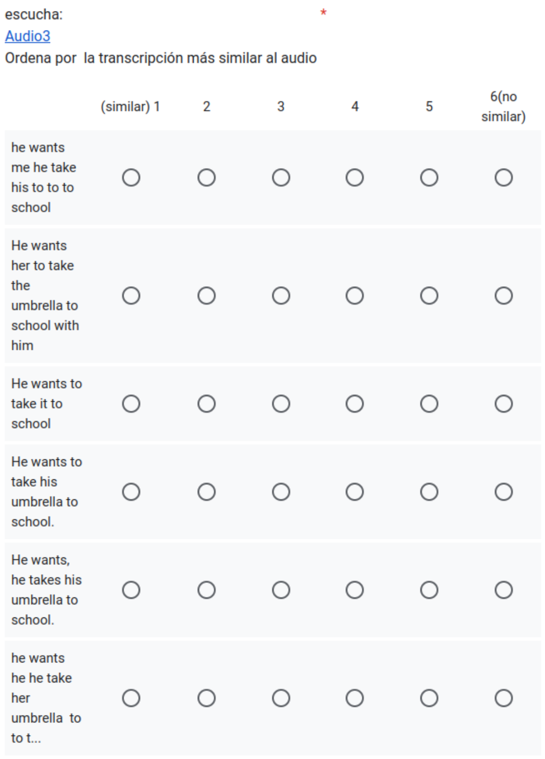
Regarding future work, it is suggested that the possibility of expanding the datasets through collaboration with organizations serving people with disabilities be considered. This could provide a more extensive corpus and enrich the fine-tuning of the systems with a greater diversity of data. Additionally, the incorporation of gestural information from speakers to assess its impact on interpreting user intent could be explored, opening up new avenues for research in this field.

APPENDIX A
Human evaluation survey

In the following appendix, a more detailed presentation will be provided regarding the type of survey questions used for the human evaluation of the different systems that combined transcription with visual context information.

As explained earlier in Chapter 9, the survey comprises 7 audio clips, with 2 questions for each audio. The first of these questions, as depicted in Figure A.1, requests the classification of systems based on the most accurate word-for-word transcription. Regarding the possible options for classification, these correspond to transcriptions obtained from the systems under evaluation: GPT-3.5 with image description, the same with key visual elements as visual context, LaMDA with description, as well as its version with image elements, the fine-tuned Whisper transcription, and the original transcription.

It is worth noting that in some questions, instead of six options for classification, there are fewer due to instances where multiple systems yield the same transcription.

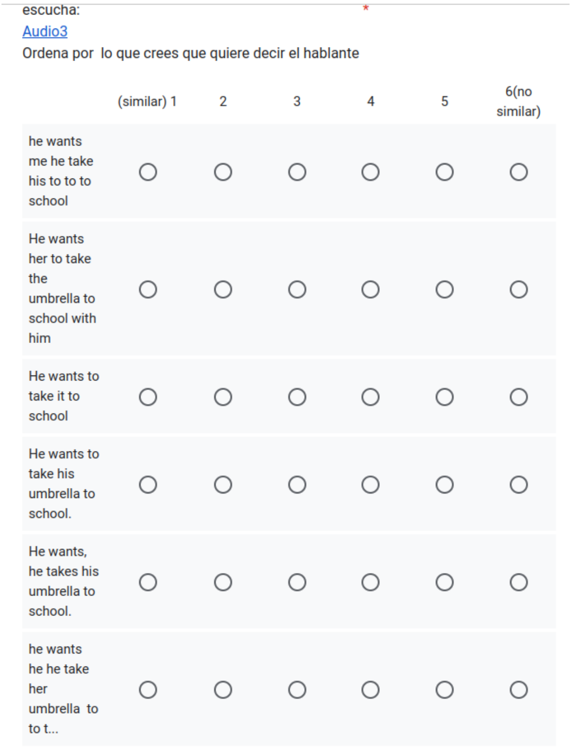


escucha: *
Audio3
Ordena por la transcripción más similar al audio

	(similar) 1	2	3	4	5	6(no similar)
he wants me he take his to to to school	○	○	○	○	○	○
He wants her to take the umbrella to school with him	○	○	○	○	○	○
He wants to take it to school	○	○	○	○	○	○
He wants to take his umbrella to school.	○	○	○	○	○	○
He wants, he takes his umbrella to school.	○	○	○	○	○	○
he wants he he take her umbrella to to t...	○	○	○	○	○	○

Figure A.1: The first type of question in the survey conducted regarding the best transcription of the presented audio.

As we can see in Figure A.2, the second question corresponding to each audio follows a similar structure and involves the same systems. However, in this case, respondents are asked to rank them based on what they believe the speaker intends to convey.



escucha: *
Audio3
Ordena por lo que crees que quiere decir el hablante

	(similar) 1	2	3	4	5	6(no similar)
he wants me he take his to to to school	○	○	○	○	○	○
He wants her to take the umbrella to school with him	○	○	○	○	○	○
He wants to take it to school	○	○	○	○	○	○
He wants to take his umbrella to school.	○	○	○	○	○	○
He wants, he takes his umbrella to school.	○	○	○	○	○	○
he wants he he take her umbrella to to t...	○	○	○	○	○	○

Figure A.2: The second type of question in the survey conducted concerning the best interpretation of the speaker's intention.

To access the audio, participants were required to click on the provided link, which would open a new tab featuring an audio player for playback.

The primary challenge encountered by survey participants was that, as individuals with pronunciation difficulties and considering the audio language was English, a proficient level of English proficiency was expected for evaluation. Nevertheless, some participants faced challenges in transcribing certain audio due to the speaker's underlying medical condition, which added complexity to the transcription task.

The comprehensive survey can be accessed via the following link: `https://forms.gle/zDwVJXbwZrb7PYr66`.

The image that is input into the VLM to obtain transcription hypotheses for evaluating the different systems in the survey can be seen in Figure A.3.

Regarding the prompts employed to introduce input into the LLM for transcription generation during the survey, we employed two distinct prompts.

In the first prompt, the key image elements obtained by the VLM are presented alongside the phrase transcribed by the ASR model, using the following prompt:

"Given these keywords about the user environment: 'cartoon', 'woman', 'child', 'umbrella', and 'rain' and given these transcriptions of the sentences said by an individual with aphasia: 'he wants me he take his to to to school ' you are an expert in aphasia transcription, return the same transcriptions of the individual modifying if you see any mistake in the transcription only if you are sure if not just return the same transcription sentence".

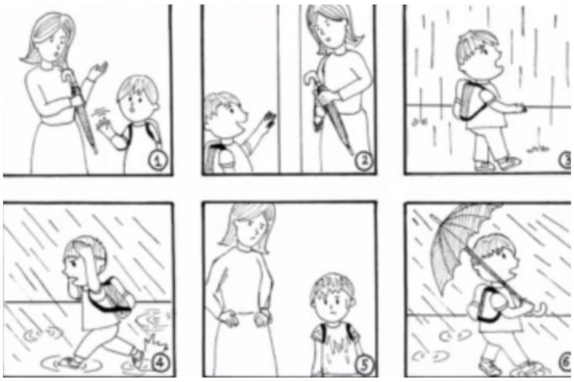


Figure A.3: The image input into the VLM LLaVAR to obtain transcriptions of the phrases in the survey conducted for human evaluation.

With regard to the second prompt, it includes a description of the image obtained by the VLM along with the transcription of the audio obtained by the ASR model, as shown below.

"Given this description about an image with a story told in 6 different cartoons the user is looking at and trying to describe: 'The image is a black and white cartoon of a woman and a child. The woman is holding an umbrella and a pencil, while the child is wearing a backpack. They appear to be interacting with each other, possibly sharing a moment or engaging in an activity. The image is a black and white cartoon depicting a woman and a child standing near each other. The woman is holding an umbrella, while the child appears to be reaching out to her or holding their hand out towards the umbrella. The woman seems to be interacting with the child, possibly teaching them about the umbrella or sharing a moment together. The image is a cartoon of a boy standing outside in the rain The image shows a cartoon of a boy with a backpack walking in the rain. He is wearing a backpack and holding an umbrella to protect himself from the rain. The boy appears to be looking up, possibly observing the raindrops or listening to the sound of the rain In the image, there is a woman and a young boy standing next to each other. The woman is on the left side, and the boy is on the right. The woman is taller than the boy and appears to have her arms crossed. The boy is looking downwards, and there is a backpack near him. The image seems to have a white background. The image depicts a cartoon boy with a backpack walking in the rain while holding an umbrella to protect himself from the downpour.', and given these transcriptions of the sentences said by an individual with aphasia: 'he wants me he take his to to to school', you are an expert in aphasia transcription, return the same transcriptions of the individual modifying if you see any mistake in the transcription only if you are sure if not just return the same transcription sentence "

Bibliography

[1] Falcon LLM — falconllm.tii.ae. https://falconllm.tii.ae/. [Accessed 07-09-2023].

[2] VOSK Models — alphacephei.com. https://alphacephei.com/vosk/models. [Accessed 03-09-2023].

[3] Vaishnavi Agrawal, Shariva Dhekane, Neha Tuniya, and Vibha Vyas. Image caption generator using attention mechanism. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2021.

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[5] Nenny Anggraini, Angga Kurniawan, Luh Kesuma Wardhani, and Nashrul Hakiem. Speech recognition application for the speech impaired using the android-based google cloud speech api. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 16(6):2733–2739, 2018.

[6] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

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

[8] Prof. S. S. Bhabad and Mrs. Anjali R. Patil. Communication aid for people with severe speech disability. 2018.

[9] Shaan Bijwadia, Shuo-yiin Chang, Bo Li, Tara Sainath, Chao Zhang, and Yanzhang He. Unified end-to-end speech recognition and endpointing for fast and efficient speech systems. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 310–316. IEEE, 2023.

[10] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[11] Cécile Cauquil-Michon, Constance Flamand-Roze, and Christian Denier. Borderzone strokes and transcortical aphasia. *Current neurology and neuroscience reports*, 11:570–577, 2011.

[12] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[13] Sang Keun Choe, Quanyang Lu, Vikas Raunak, Yi Xu, and Florian Metze. On leveraging visual modality for speech recognition error correction, 2019.

[14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[15] Yan Chu, Xiao Yue, Lei Yu, Mikhailov Sergei, and Zhengkui Wang. Automatic image captioning based on resnet50 and lstm with soft attention. *Wireless Communications and Mobile Computing*, 2020:1–7, 2020.

[16] Antonio R Damasio. Aphasia. *New England Journal of Medicine*, 326(8):531–539, 1992.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Johanes Effendi, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Listening while speaking and visualizing: Improving asr through multimodal chain. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 471–478. IEEE, 2019.

[19] Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. A bibliometric review of large language models research from 2017 to 2023. *arXiv preprint arXiv:2304.02020*, 2023.

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

[21] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[22] Jordan R Green, Robert L MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A Ladewig, Jimmy Tobin, Michael P Brenner, et al. Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. In *Interspeech*, pages 4778–4782, 2021.

[23] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[24] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[25] Biing-Hwang Juang and Lawrence R Rabiner. Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, 1:67, 2005.

[26] Arbana Kadriu and Amarildo Rista. Automatic speech recognition: A comprehensive survey. *Seeu Review*, 15(2):86–112, 2020.

[27] June-Woo Kim, Ho-Young Jung, et al. End-to-end speech recognition models using limited training data. *Phonetics and Speech Sciences*, 12(4):63–71, 2020.

[28] Suraj Kothawade, Anmol Mekala, D Chandra Sekhara Hetha Havya, Mayank Kothyari, Rishabh Iyer, Ganesh Ramakrishnan, and Preethi Jyothi. Ditto: Data-efficient and fair targeted subset selection for asr accent adaptation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5810–5822, 2023.

[29] Oleksii Kuchaiev, Boris Ginsburg, Igor Gitman, Vitaly Lavrukhin, Carl Case, and Paulius Micikevicius. Openseq2seq: extensible toolkit for distributed and mixed precision training of sequence-to-sequence models. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 41–46, 2018.

[30] Vanya Bannihatti Kumar, Shanbo Cheng, Ningxin Peng, and Yuchen Zhang. Visual information matters for asr error correction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[31] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[32] Jinyu Li et al. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022.

[33] Jinyu Li, Yu Wu, Yashesh Gaur, Chengyi Wang, Rui Zhao, and Shujie Liu. On the comparison of popular end-to-end models for large scale speech recognition. *arXiv preprint arXiv:2005.14327*, 2020.

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[36] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. A frustratingly simple approach for end-to-end image captioning. *arXiv preprint arXiv:2201.12723*, 2022.

[37] Brian MacWhinney. The talkbank project. *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, pages 163–180, 2007.

[38] Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307, 2011.

[39] Iosif Mporas, Todor Ganchev, Mihalis Siafarikas, and Nikos Fakotakis. Comparison of speech features on the speech recognition task. *Journal of Computer Science*, 3(8):608–616, 2007.

[40] OpenAI. Gpt-4 technical report, 2023.

[41] D. O'Shaughnessy. Interacting with computers by voice: automatic speech recognition and synthesis. *Proceedings of the IEEE*, 91(9):1272–1305, 2003.

[42] Youssef Oualil and Dietrich Klakow. Image-sensitive language modeling for automatic speech recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[43] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

[44] Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. End-to-end speech recognition: A survey. *arXiv preprint arXiv:2303.03329*, 2023.

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

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[47] Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 193–199. IEEE, 2017.

[48] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

[49] AM Redondo Romero and J Lorente Aledo. Trastornos del lenguaje. *Pediatría Integral, VIII (8), 675*, 691, 2006.

[50] Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, et al. Personalizing asr for dysarthric and accented speech with limited data. *arXiv preprint arXiv:1907.13511*, 2019.

[51] Rohini K Srihari. Automatic indexing and content-based retrieval of captioned images. *Computer*, 28(9):49–56, 1995.

[52] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[53] Katrin Tomanek, Vicky Zayats, Dirk Padfield, Kara Vaillancourt, and Fadi Biadsy. Residual adapters for parameter-efficient asr adaptation to atypical and accented speech. *arXiv preprint arXiv:2109.06952*, 2021.

[54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[55] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

[57] Chaoyang Wang, Ziwei Zhou, and Liang Xu. An integrative review of image captioning research. In *journal of physics: conference series*, volume 1748, page 042060. IOP Publishing, 2021.

[58] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*, 2018.

[59] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*, 2023.

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

[61] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

[62] Luowei Zhou, Chenliang Xu, Parker Koch, and Jason J Corso. Watch what you just said: Image captioning with text-conditional attention. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 305–313, 2017.

[63] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.