

Improving Aphasic Communication Using Multimodal AI Systems

Isabel Ferri-Molla , Jordi Linares-Pellicer , Juan Izquierdo-Domenech  *

Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (UPV), Camí de Vera, s/n, 46022, Valencia (Spain)

* Corresponding author: isfermol@upv.es (I. Ferri-Molla), jorlipel@upv.es (J. Linares-Pellicer), juaizdom@upv.es (J. Izquierdo-Domenech).

Received 9 August 2024 | Accepted 18 December 2025 | Early Access 9 March 2026



ABSTRACT

Aphasia, often resulting from brain injuries, significantly impairs individuals' language abilities, creating substantial challenges for verbal communication. Existing assistive technologies frequently fall short in addressing these specialised communication needs, underscoring the urgent demand for adaptive, intelligent support systems. This research proposes a dual approach: an Automatic Speech Recognition (ASR) module fine-tuned on aphasic speech, and a multimodal component that integrates visual context to infer the speaker's intended meaning. The ASR system leverages fine-tuned versions of Whisper and Wav2Vec 2.0 on data from the AphasiaBank corpus. Results show a notable reduction in Word Error Rate (WER) when comparing base pre-trained ASR models with their finetuned versions, decreasing from 70.36% to 31.53% in a context-independent setting, and from 61.25% to 35.60% in a speaker-independent evaluation, demonstrating robustness across different scenarios. In contrast to the ASR module, the goal of the multimodal component is not to produce a literal word-by-word transcription, but rather to reconstruct the speaker's communicative intent using contextual information. To evaluate this capability, we conducted a human study assessing the system's ability to interpret what the speaker truly meant. The results confirmed that outputs combining visual cues with language model reasoning more reliably captured communicative intent than audio-only transcriptions.

KEYWORDS

Aphasia, ASR, HCI, Image Captioning, Multimodality.

DOI: 10.9781/ijimai.2026.2215

I. INTRODUCTION

SPEECH recognition, the conversion of spoken language into written text, has witnessed remarkable advancements in recent years, driven by the increasing use of spoken language as a way of interacting with diverse systems and devices. Automatic Speech Recognition (ASR) systems, integral to various domains like chatbots, voice assistants, and translation systems, have progressed from traditional methods like Hidden Markov Models [1] to modern Deep Learning (DL) techniques like transformers [2]. This progression has given rise to the emergence of several renowned solutions, including Whisper [3], Wav2Vec 2.0 [4], and Seamless [5].

The expansion in ASR technology has underscored the need for enhanced accessibility. Despite the significant progress in ASR research and development, existing voice recognition systems encounter difficulties in accurately transcribing speech from specific user groups with pronunciation complexities. This limitation results in frustration and restricted technology adoption within these communities.

Addressing the complexity, inaccuracy, and variations in pronunciation, recent studies have shown promising results through fine-tuning advanced models for specific user domains. Notable examples highlight the potential adaptation of speech recognition systems for individuals with pronunciation difficulties, sparking optimism about facilitating effective communication for these users. However, due to the complexity of the domain, even after adapting and fine-tuning the model, it is common to still face a high rate of incomplete information. This means that, even if the transcription of what the user says is improved, the speaker may not be able to precisely express what he or she wants to convey.

This study seeks to enhance the communicative capabilities of individuals with speech impairments by integrating ASR and image description technologies within Large Language Models (LLMs). The focus is on developing systems able not only of transcribing speech from individuals with pronunciation challenges but also of extracting visual context through image description systems. This integration is essential for compensating speech imprecision and leveraging visual cues, thereby improving the accuracy of interpretation for users with pronunciation difficulties. Effectiveness is evaluated through a

Please cite this article as:

I. Ferri-Molla, J. Linares-Pellicer, J. Izquierdo-Domenech. Improving Aphasic Communication Using Multimodal AI Systems, *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 9, no. 7, pp. 67-77, 2026, <http://doi.org/10.9781/ijimai.2026.2215>

combination of quantitative and qualitative methods, including Word Error Rate (WER) during ASR fine-tuning and a human-centered evaluation to assess semantic understanding in multimodal scenarios. The corpus used, offers rich, structured recordings of individuals with aphasia enabling detailed analysis of both transcription accuracy and communicative intent. The objective is to demonstrate that such a system can significantly improve the quality of interaction and enable better societal integration for affected individuals.

The sub-goals of the project involve data collection and cleaning, partitioning strategies, model selection, fine-tuning and adjusting weights, and evaluating the impact of contextual image descriptions on speech comprehension. The ultimate aim is to develop a multimodal system that generates messages aligning more closely with the speaker's intention, thereby fostering improved communication for individuals facing challenges in speech articulation in the real world.

Contributions of this work can be summarized as follows:

- **Development of a Fine-Tuned ASR Approach:** The study fine-tunes state-of-the-art pre-trained ASR models, such as Whisper, to the specific characteristics of aphasic speech, resulting in significant reductions in WER compared to their non-adapted counterparts.
- **Speaker-independent and Context-independent partitioning:** Two complementary data partitioning strategies are explored, based on speaker identity and thematic context, demonstrating the robustness and generalization capability of the fine-tuned ASR models across different speech patterns.
- **Multimodal integration with visual context:** Beyond transcription, the system incorporates image captioning techniques to provide visual environmental cues. These cues are combined with ASR outputs and processed by an LLM to better capture the intended meaning behind speech utterances, especially when verbal expression is incomplete or ambiguous.
- **Human-Centered Evaluation:** A survey-based user study assesses the system's effectiveness both in terms of literal transcription accuracy and semantic alignment with the speaker's intent. Results show that while ASR fine-tuning improves word-level accuracy, the multimodal system provides better semantic understanding.
- **Practical implications for Aphasia:** By combining speech recognition accuracy, achieved through fine-tuned ASR models on aphasic speech, with semantic interpretation grounded in visual context, the proposed system represents a significant step toward inclusive communication technologies. This multimodal integration offers tangible improvements for individuals with aphasia, enabling more accurate understanding of their intended messages and facilitating interaction with both digital systems and broader society.

II. RELATED WORK

Since the shift in the ASR field from purely statistical models to incorporating neural networks, there has been substantial growth. The most significant change was the replacement of gaussian mixtures in the acoustic model with Deep Neural Networks (DNN) [6].

Traditional ASR systems were constructed using discrete elements, specifically an acoustic model, a language model, and a vocabulary. These components were fine-tuned and trained individually. However, most of the current state-of-the-art models in the ASR field lie nowadays in end-to-end models [7].

These models have revolutionized the field by employing a single neural network to directly transcribe audio signals into text. This end-

to-end approach offers several advantages, including the optimization of a single objective function and a more compact system design due to the reduced number of intermediate components. End-to-end architectures have been widely explored in automatic speech recognition tasks, demonstrating their effectiveness across different scenarios [8], [9].

Previous research in speech recognition has provided valuable insights into improving recognition accuracy and robustness in noisy environments. In particular, the use of Simple Recurrent Units (SRUs) combined with attention mechanisms has been shown to enhance speech intelligibility and recognition performance [10].

Regarding voice recognition for users experiencing pronunciation difficulties, prior work has focused on adapting speech recognition technologies to meet the specific needs of these populations. Speaker-dependent keyword detection approaches have been shown to be effective for individuals with speech disorders, enabling more personalized ASR systems. In this context, convolutional neural network-based solutions have been explored to tailor recognition models to users with dysarthria, demonstrating the importance of customization for improved performance [11]. Although dysarthria differs from aphasia and these approaches rely on convolutional architectures rather than the transformer-based models investigated in this work, the underlying motivation to adapt ASR systems for users with special needs remains highly relevant.

Recent research on aphasia has explored the use of computational models to assess the severity of language impairments, including approaches aimed at estimating the aphasia quotient of patients [12]. In parallel, broader analyses have examined the role of Artificial Intelligence in aphasia-related applications, highlighting current methodologies and emerging trends in the field [13].

Due to the large scale of many state-of-the-art models, finetuning has emerged as a promising and widely adopted technique. This approach builds upon pre-trained models and adapts them to narrower, task-specific domains, enabling effective specialization without training from scratch [14].

In the context of fine-tuning strategies aimed at customizing speech recognition systems for individuals with pronunciation difficulties, recent research has explored different personalization approaches to improve recognition performance. Both speaker-independent and personalized ASR models have been evaluated for users with speech disorders, including commercial systems and end-to-end architectures based on Recurrent Neural Network Transducer (RNN-T) models [15], [16].

Experimental results have shown that adapting ASR models to individual users can lead to substantial reductions in WER, particularly for speakers with moderate to severe speech disorders. In some cases, personalization has achieved WER reductions of up to 80%, demonstrating the effectiveness of fine-tuning for these populations [15]. However, these approaches typically rely on training a distinct personalized model for each speaker, which may limit their scalability in real-world applications. Similar benefits of fine-tuning have been reported when adapting ASR systems for individuals with amyotrophic lateral sclerosis [17].

Beyond personalization, speech is a dynamic and continuous process influenced not only by the linguistic content it conveys but also by the context in which the message is delivered. From this perspective, incorporating complementary modalities, such as visual context through image description systems, represents a promising direction to further enhance speech comprehension.

The integration of visual information has been explored as a strategy to improve the robustness of automatic speech recognition systems. One approach investigates the use of image captions generated

by captioning models as auxiliary context to correct ASR outputs, demonstrating the potential of multimodal cues for transcription refinement [18].

Two complementary strategies have been proposed to achieve this objective. One approach combines visual and textual representations by fusing embeddings derived from image features and ASR transcriptions, which are subsequently processed by a decoder to correct transcription errors using visual context. An alternative strategy introduces image captions and ASR transcriptions as separate inputs, using both as prompts for an encoder–decoder model so that visual descriptions can guide the correction of transcription inaccuracies [18].

While these methods highlight the effectiveness of visual context for improving ASR outputs, the evaluation is limited to general spoken English audio and corresponding transcriptions. As a result, the reported findings may not directly generalize to scenarios involving users with speech impairments, where pronunciation patterns and error characteristics differ substantially.

Beyond caption-based approaches, multimodal pipelines have addressed combinations of speech-to-text, text-to-speech, image-to-text, and text-to-image tasks, illustrating the broader applicability of multimodal processing frameworks to ASR-related problems [19]. Additionally, audio–visual integration strategies based on handcrafted visual descriptors and feature selection techniques have been shown to improve speech recognition performance under noisy conditions, further supporting the value of multimodal information for robust ASR [20].

In contrast, an alternative strategy focuses on post-processing ASR outputs through error correction models to enhance recognition performance. This approach generates an n-best list of candidate transcriptions, which is subsequently rescored using a combination of LLMs and visual–semantic joint embedding techniques to select the most accurate output [21].

Recent research has increasingly explored multimodal approaches to enhance ASR performance for individuals with speech impairments. For instance, AV-HuBERT has been effectively applied to reconstruct intelligible speech from dysarthric input by integrating audio and facial video features [22]. Similarly, Yu et al. [23] introduced a multi-stage fusion strategy combining acoustic and visual cues, which led to significant reductions in WER using facial images. Beyond technical advances, several studies have begun to address the real-world deployment of such technologies, highlighting both their potential and limitations. Howarth et al. [24] conducted participatory evaluations of the Voiceitt app with users with dysarthria, revealing important usability challenges and the need for customization in voice-based interfaces. Ayoka et al. [25] examined the use of Google Project Relate with 15 users in Ghana, reporting critical accessibility barriers related to accent variation, language mismatch, and severity of impairment. These works reflect growing awareness of generalization issues and underscore the diversity of needs among users with speech impairments.

While most existing ASR and AAC solutions rely on predefined phrases or require significant manual input, thereby limiting spontaneous and semantically rich communication, the approach proposed in this study combines personalized ASR, visual context interpretation, and LLMs to better capture user intent. Although previous multimodal systems have typically incorporated visual information to rectify transcription errors, our goal extends well beyond improving raw accuracy. In the context of aphasia, where speech may be incomplete, ambiguous, or semantically fragmented, precise transcription alone is often insufficient for effective interaction. Instead, our system leverages visual grounding and

contextual language understanding to infer what the user truly intends to express. This shift, from merely transcribing words to actively interpreting communicative intent, represents a critical advancement toward inclusive, adaptive technologies for individuals with expressive language impairments.

III. DATA SELECTION AND STRUCTURE

In order to adapt previously trained models to the specific characteristics of individuals facing pronunciation difficulties, it is imperative to possess a suitable corpus covering a wide collection of speech recordings from people affected by such problems, accompanied by their corresponding transcriptions.

During the dataset collection, the Talkbank database [26] was identified and utilized. Talkbank, supported by a collaborative network of hundreds of contributors, encompasses repositories spanning 14 distinct research areas in more than 34 languages, focusing particularly on spoken language.

Between the several research areas data included in talkbank, this research works with Aphasia Bank [27], a database with recordings and transcriptions in several languages conceived for the study of communications in people who suffer from aphasia.

Regarding the data from the Aphasia Bank, datasets for both Spanish and English speakers have been selected with the intention of fine-tuning two different models, one in Spanish and the other in English. In the fine-tuning processes for both languages, a similar approach has been followed.

The selected Spanish dataset comprises four videos, each approximately 40 minutes long, while the English dataset selected for this project 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 a story.

The dataset for training the proposed system comprises audio recordings in a dialogue format with corresponding transcriptions in .cha format. Each sentence is enriched with metadata, including speaker information (patient or interviewer), language modifications, timestamps, and POS tagging.

In the initial data processing phase, emphasis was placed on sentence transcriptions, disregarding other metadata. Specifically, only sentences spoken by patients with aphasia were used for training, omitting any utterances from the interviewer. This approach effectively halved the total audio duration per video, resulting in approximately 15–20 minutes per recording.

The alignment between audio and transcription was performed using the manually annotated timestamps from the AphasiaBank corpus. These timestamps define the exact start and end time of each utterance and were precise enough to avoid the need for additional forced alignment tools. Nonetheless, utterances with unclear timing, overlapping speech, or audio quality issues were manually reviewed or excluded to ensure clean, well-aligned data.

However, it is worth noting that the videos from the AphasiaBank corpus were recorded in a controlled environment. The recordings took place in a classroom where patients performed structured exercises, which ensured consistent audio quality across the recordings and extracted segments.

To preprocess the data, a script was used to extract the corresponding video segments based on the timestamps associated with the selected

patient sentences, which were identified through the speaker labels in the transcriptions, namely, only those marked as patient were included. A CSV file was generated containing the clipped segments, with audio exclusively from the person with aphasia, along with their transcriptions. This process was applied uniformly to all videos.

To prepare a specialized dataset for training, scripts using regular expressions were employed to filter and adapt the transcriptions, ensuring they conformed to the format required by the target models. This meticulous process ensured that the models were trained effectively using relevant speech data from patients, addressing specific challenges associated with aphasia.

To structure the dataset for training and evaluation, the final CSV was randomly split into training (75%), validation (10%), and test (15%) sets. This partitioning ensured proper separation of data for model fine-tuning and performance assessment.

IV. APPROACH AND METHODOLOGY

The proposed system follows a two-stage architecture. In the first stage, several pre-trained ASR models are fine-tuned using the aphasia-specific dataset described in Section III. This adaptation enables the models to better capture the speech patterns of individuals with aphasia, significantly improving transcription accuracy.

However, this enhancement alone is not always sufficient for effective communication. Some individuals with specific types of aphasia face expressive limitations that go beyond pronunciation errors, such as difficulty retrieving the correct words or forming coherent phrases. In such cases, even a perfectly accurate word-for-word transcription may fail to convey the speaker's true intent.

To address this limitation, the second stage of the system introduces a multimodal strategy designed to enrich the transcription with contextual visual information. By combining the audio transcription with images of the user's environment and processing both inputs through an LLM, the system can infer the intended meaning more accurately, even when the spoken message is incomplete, imprecise, or ambiguous.

The following section describes the components and methodology of this integrated system in greater detail.

In the multimodal system proposed in this work, three key components are considered: an ASR model, an image description model, and a LLM. Fig. 1 provides a visual representation of the entire process.

The first stage involves processing audio recordings of users and transcribing them using the ASR model previously fine-tuned on aphasic speech. This adapted model is employed to generate the most accurate possible transcriptions of the users' utterances, taking into account the specific speech patterns associated with aphasia.

Simultaneously, images capturing the user's visual context are processed using a Vision-Language Model (VLM), which interprets the visual input and generates a natural language description of the user's surroundings. Two prompting strategies are explored to guide the VLM output. The first strategy requests a natural language description of the scene, resulting in a detailed caption that captures the overall context. In the second strategy, the prompt explicitly requests a "list of the main visible objects," leading the model to generate a concise set of keywords (e.g., "bed, nightstand, window, lamp"). These keywords are parsed from the model's textual response using rule-based postprocessing and represent the most salient elements in the image. These two types of outputs, full-sentence captions and keyword lists, are later compared in terms of their usefulness for enhancing the transcription through multimodal integration.

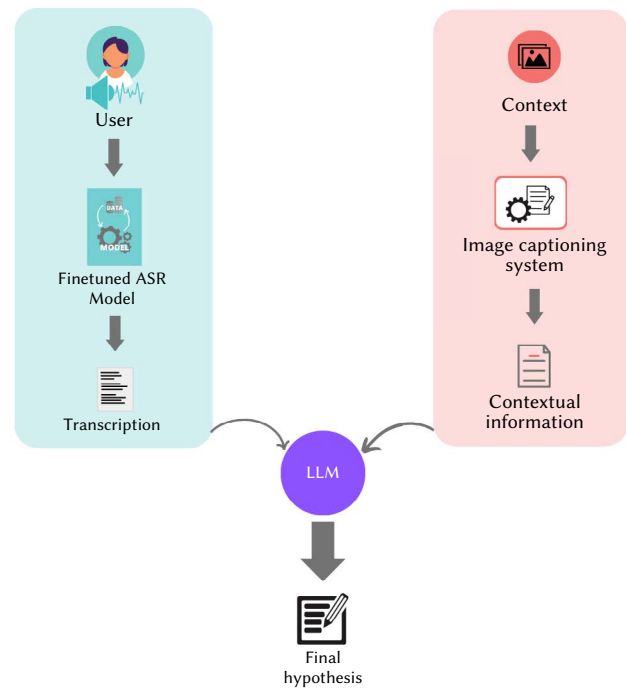


Fig. 1. Diagram of the complete multimodal system developed in the presented work.

In the final stage, the audio transcript generated by the ASR model fine-tuned on aphasic speech, together with the visual context (either the caption or the keyword list), are combined into a single structured prompt and passed as input to a LLM. This approach leverages the capabilities of instruction-tuned LLMs to process language-based prompts without requiring additional multimodal training. Recent studies have shown that using prompt-based integration of textual and visual inputs can be more effective and computationally efficient than training joint multimodal embeddings, especially when working with large pre-trained models [28], [29]. The LLM uses the combined prompt to refine or complete the transcription, leveraging the visual context to better infer the speaker's intended message when the original utterance is incomplete or ambiguous.

After completing the solution development, an analysis of the results obtained was conducted, focusing on two distinct approaches to evaluate effectiveness. Firstly, an assessment has been performed using only the fine-tuned ASR model. This phase involves comparing different ASR models and experimenting with various hyperparameters to optimize the fine-tuning process and achieve optimal results.

On the other hand, the previously described complete system has been assessed. This system integrates the interpretation of environmental descriptions or keywords generated by the VLM. It is responsible for refining transcriptions suggested by the ASR model based on the contextual information derived from the images.

V. RESULTS AND EVALUATION

The evaluation of the proposed system involves a series of experiments using different ASR architectures. In the first phase, multiple pre-trained models of varying sizes are fine-tuned to the domain of aphasic speech in order to assess whether this adaptation leads to measurable improvements over baseline systems. All experiments were implemented in Python, using the TensorFlow library for model training and data processing¹.

¹ The code used in this study is publicly available at <https://github.com/Vertextlit-VRAIN/aphasia>

The process begins with the retrieval of data from Aphasia Bank. Transcriptions are then refined to extract only the relevant segments, and video files are converted into audio. Sentence-level snippets are aligned with their corresponding transcripts based on available timestamps. The resulting data is stored in a CSV file containing paths to segmented audio files and corresponding transcriptions. Audio from interviewers is excluded to ensure that only speech from individuals with aphasia is used.

The CSV is shuffled and partitioned into "train" (75%), "test" (15%), and "validation" (10%) data frames. Each subset undergoes tokenization, preprocessing, and feature extraction. The pre-trained ASR model is then fine-tuned using this data to better accommodate the characteristics of aphasic speech.

To evaluate the effectiveness of the finetuned ASR model, the WER metric, based on transcribed sentences from the validation set and the original transcriptions of the sentences, is used.

The initial data partitioning is approached from two different perspectives: a context-independent approach and a speaker-independent approach. In the context-independent approach, the train, validation, and test splits are based on thematic or contextual content, allowing the examination of the impact of fine-tuning on transcriptions across various topics. On the other hand, the speaker-independent approach divides the data solely based on the speaker's identity, disregarding the underlying context or subject matter.

A. Context Independent Tests

As previously described, each video in the dataset includes a segment where participants describe various images. This visually grounded portion of the recordings provides an ideal scenario for evaluating the impact of contextual information on transcription accuracy. In the context-independent testing phase, a specific segment of videos is extracted, focusing on recordings of participants describing specific images. This segment is removed from all videos respectively in the training and testing sets, and used exclusively in the validation process, ensuring exposure for validation to data unseen in training.

The decision to reserve the audio segment where images are described for validation is strategic, allowing for the integration of visual context in the subsequent phase.

Once data is prepared, partitioned, and tokenized, and training and test datasets are established, the crucial stage is defining training parameters for fine-tuning. Hyperparameters such as batch size, learning rate, warmup steps, max steps, save and evaluation frequency, or maximum length of generated sequences are determined.

Comprehensive experiments were conducted to identify optimal parameter values for the model, with a key focus on the batch size during training. Considering constraints related to data size and system capacity, the batch size used was 8. Trials indicated that a larger batch size generally led to improved outcomes.

Hyperparameter tuning was initially performed using the "whisper-small" model and the English subset of the AphasiaBank dataset.

To determine which hyperparameters performed better in each case, the WER was used as a performance metric in the finetuning process. WER is widely adopted in ASR to evaluate how closely the model's output matches a reference transcription. It is formally defined as shown in Equation 1.

$$\text{WER} = \frac{S + D + I}{N} \quad (1)$$

Where S is the number of substitutions (when one word is incorrectly replaced by another), D is the number of deletions (words present in the reference but missing in the hypothesis), I is the number of insertions (extra words added by the model), and N is the total

number of words in the reference transcription. Each of these error types contributes equally to the total WER.

For example, if the model transcribes 'the dog run' instead of the reference 'the dog runs fast', it may incur one substitution ('run' for 'runs'), one deletion ('fast' is omitted), and no insertions. A lower WER reflects a higher level of transcription accuracy and better alignment with the ground truth.

Although the WER is a standard metric for assessing ASR performance, its limitations are particularly relevant in the context of aphasic speech. WER is used in this study exclusively during the fine-tuning stage of the ASR model, where the goal is to maximise transcription fidelity to the original audio. In this phase, preprocessing steps are carefully designed to minimise the effect of noise, misalignment, or unintelligible speech, thereby ensuring that WER accurately reflects model improvement. While aphasic speech may contain disfluencies or incomplete structures, WER still provides a meaningful measure of whether the model has better learned to reproduce what was actually spoken. It is worth noting that the same limitations of WER apply equally to pre-trained and fine-tuned models, and therefore do not invalidate the relative improvements observed.

However, in the second stage of the system, when visual context is incorporated through multimodal integration, the objective shifts from literal transcription to capturing the speaker's intended meaning. In this scenario, WER becomes inadequate, as it does not account for semantic equivalence, context dependence, or communication effectiveness. For this reason, the final system is evaluated through a human user study, in which participants judge whether the output better reflects what the speaker likely meant to say. This evaluation provides a more meaningful assessment for real-world assistive applications.

The detailed results for different hyperparameter values are presented in Table I as well as in Fig. 2.

TABLE I. ADAPTATION OF HYPERPARAMETERS IN THE WHISPER-SMALL MODEL USING THE APHASIA ENGLISH DATASET AND A CONTEXT-INDEPENDENT APPROACH

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

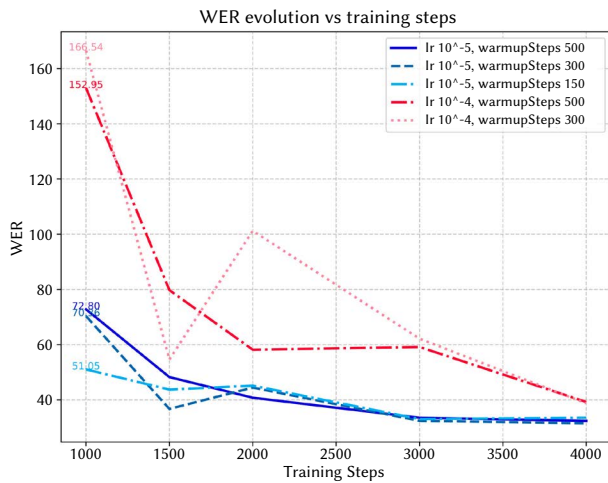


Fig. 2. WER evolution depending on the selected hyperparameters during the fine-tuning process of the Whisper-Small model.

As outlined in Table I and Fig. 2, 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} . Interestingly, with a learning rate of 10^{-5} , the WER evolution from step 3000 onwards appears to stabilize, indicating a more modest improvement. These optimized values were subsequently employed in the next fine-tuning process.

In the experimentation process concerning the fine-tuning of various ASR systems, three sizes of Whisper models (small, base, and tiny) have been selected in addition to the wav2vec 2.0 model. For each of these models, the fine-tuning process previously elucidated has been executed, utilizing the data acquired from the English dataset in AphasiaBank.

The WER results obtained from a validation set, which includes data representing a distinct audio theme not present in the training data, are presented for both the base models (models without the fine-tuning process) and the fine-tuned models in Table II. It is imperative to note that these results stem from data partitioned following the content-independent approach.

TABLE II. ACHIEVED WER IN THE EXPERIMENTS WITH THE ENGLISH DATASET USING THE CONTEXT-INDEPENDENT APPROACH

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

As depicted in Table II, it is evident that the Whisper-small model experienced a substantial improvement, reducing the WER from 70.36% to 31.53%. This signifies a remarkable relative improvement of 55.19%. Similarly, the Whisper-base model reflected a relative improvement of 39.53%. The huge improvement in model performance after fine-tuning is also influenced by the specificity of the aphasia data and its difference from the data presumably used in training the base pre-trained models.

In addition to the training process for various models using the English dataset, preliminary tests were conducted in Spanish with a more limited dataset. These tests involved a direct comparison between the pre-trained "small" Whisper model, the pre-trained ESPnet2 model [30], and the fine-tuned Whisper model on the Spanish dataset.

For the fine-tuning of the Whisper model on the Spanish language dataset, specific hyperparameters were employed, including a learning rate of 10^{-5} , a batch size of 8, an initial warm-up step count of 500, and a total of 3000 training steps.

TABLE III. ACHIEVED WER IN THE EXPERIMENTS CONDUCTED WITH THE SPANISH DATASET

Model	WER(%)
ESPnet2	182.33
Whisper-small	121.57
fine-tuned Whisper-small	36.34

As depicted in Table III, the tests conducted in Spanish demonstrated a significant performance gap between models fine-tuned with contextual data for individuals with aphasia and baseline pre-trained models. Specifically, for the Whisper model, the WER indicated a substantial relative improvement of 70.11%.

In context-independent testing, WER may be influenced by specific segments of image descriptions used as the validation set. This dependence on selected images introduces variations in metric outcomes based on the vocabulary specificity of the image elements. Notably, experiments, especially those involving individuals with speech difficulties, highlight the crucial role of the fine-tuning process. The observed improvement is presumed to stem from the end-to-end model component, analogous to the role played by the acoustic model in hybrid systems, enhancing the model's comprehension of user speech patterns. Consequently, a second experiment will be conducted, basing the division of training, testing, and validation sets on speakers rather than content. This approach aims to enhance the model's accuracy and usability for users with diverse speech patterns and speech difficulties.

B. Speaker Independent Tests

After testing the initial model, which utilized context-based data splitting, and obtaining previous conclusions, an alternative approach was adopted. In this experiment, the dataset was partitioned according to individual speakers, to ensure distinct speaker representation. No speaker appears in more than one subset (training, validation, and testing), thus allowing for a true assessment of the model's generalization to unseen speakers. A consistent validation set was used for all experiments, with an increased dataset size that included recordings from different speakers in the English subset of AphasiaBank. The training set comprises approximately 720 minutes, while the validation set has around 105 minutes. The training process mirrored previous experiments, and hyperparameter selection, including a learning rate of 10^{-5} , 300 warm-up steps, and a maximum of 3000 steps, followed the criteria outlined earlier.

TABLE IV. ACHIEVED WER OF THE EXPERIMENTS WITH THE ENGLISH DATASET IN SPEAKER-INDEPENDENT APPROACH

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

The Table IV illustrates a consistent trend among base models, where larger models with more parameters show better WER metrics. In all cases, the fine-tuned models outperform their non-fine-tuned counterparts in terms of WER. Whisper-small achieves the best WER in this set of tests at 35.60%, marking a 41.88% relative improvement compared to the non-fine-tuned Whisper-small model's WER of 61.25%.

It is crucial to emphasize that these results were obtained using a different partitioning approach from the context-independent set of experiments, leading to variations in the validation set and explaining the differences in results.

To contextualize our WER results within the broader field of impaired speech recognition, we briefly discuss related studies addressing similar challenges, despite differences in speech disorders and experimental conditions. WER is a widely used metric in ASR evaluation. However, its interpretation across studies must be made with caution due to significant differences in speech disorder types, dataset composition, languages, and system modalities. For instance, Chen et al. [22] applied a multimodal approach based on AV-HuBERT to reconstruct Mandarin dysarthric speech, achieving a WER reduction from 85.2% to 52.8% on average. Similarly, Yu et al. [23] evaluated a multi-stage audio-visual fusion model on English dysarthric speech, reporting WERs of 6.05% for mild, 22.8% for moderate, 30.77% for severe, and 63.98% for extremely severe cases. In contrast, the present work addresses aphasic speech, a distinct neurological condition, and explores the fine-tuning of Whisper and Wav2Vec 2.0 models on English and Spanish data from the AphasiaBank corpus. While these values are comparable to those obtained in previous studies for severe dysarthria, direct numerical comparison is not appropriate due to the aforementioned differences. Nonetheless, situating our results within this broader range contributes to understanding the WER scale in the context of speech disorders. Furthermore, while our fine-tuning strategy is designed to improve word-level transcription accuracy, the multimodal stage of our system is not intended to optimize WER, but rather to enhance semantic alignment with the speaker's communicative intent, which is an especially relevant goal in the context of aphasia, where speech is often fragmented, ambiguous, or incomplete.

Although WER provides a clear quantitative indicator of transcription accuracy, its true impact is best understood when considering real-world communication outcomes. For individuals with aphasia, even modest improvements in WER can lead to significantly more effective interactions with caregivers, family members, or assistive systems. A lower WER not only reduces misunderstandings, but also increases the likelihood that the user's intended message is successfully conveyed.

For instance, a user might attempt to say "want water" but produce a disfluent utterance. A generic ASR model might incorrectly transcribe this as "one war," leading to confusion. In contrast, a fine-tuned model trained on aphasic speech might correctly interpret it as "want water," preserving the core message. Moreover, when enhanced with visual context and processed through a language model, as will be explained in the next subsection, the final output might be "I want a glass of water," capturing not just the literal words but the intended meaning. This example illustrates how domain adaptation and multimodal integration can substantially improve user experience, reduce frustration, and build trust in voice-based assistive technologies.

Nevertheless, while WER provides valuable insights into model performance, it does not account for semantic understanding or the user's communicative intent. To address this limitation, and in order to evaluate the multimodal system presented in the next subsection, a human evaluation study was conducted to assess the system's ability to infer and convey what the speaker truly meant to say, particularly in cases of disfluent, incomplete, or ambiguous speech.

The following section details the design, methodology, and results of this evaluation, highlighting its importance in validating the system's effectiveness in real-world communicative contexts.

C. Multimodal System Test

Following the promising results obtained from the ASR model fine-tuning experiments, a new set of experiments was conducted to evaluate the efficacy of incorporating visual context to enhance translation within this particular domain. The hypothesis posits that, as certain types of aphasia may hinder the ability of patients to express themselves adequately, this could potentially lead to suboptimal interpretations through transcription alone. In such specific cases, the inclusion of visual context would be very valuable, as it has the potential to provide information for obtaining a more accurate interpretation of what the user tries to communicate.

To conduct the upcoming experiments, a multimodal system has been developed. For the ASR model employed within this system, the context-independent Whisper-small model fine-tuned with the English dataset has been selected. This decision was made due to its superior WER performance in context-independent tests. Furthermore, adopting a context-independent approach for the multimodal system aims to mitigate the risk of overfitting in the results, owing to the nature of the multimodal system and the evaluation conducted. To validate the system, it is imperative that the context associated with the images in question has not been previously encountered during the system's training phase. The meticulous data partitioning process used in the context-independent approach has been critical in achieving this, ensuring the pre-trained system has not been exposed to the precise contextual information designated for validation, thereby maintaining the integrity of the assessment of the multimodal system.

In this multimodal procedure, in addition to the audio that will be transcribed by the fine-tuned ASR model, a picture of the user's environment will also be provided as input to the system. From this image, a VLM will extract information about the visual context of the user. Two discrete methodologies will be applied in leveraging the VLM. Primarily, it will generate a list of elements portrayed within the image, and in a separate test, it will furnish an elaborate description of the image. To elicit these distinct outcomes, two different prompts were introduced to the VLM alongside the corresponding image, contingent upon whether the anticipated output is a descriptive narrative or a listing of principal elements.

The response of the VLM and the audio transcription from the fine-tuned ASR system will be combined into a prompt that is fed into a LLM, which then produces a modified transcription considering both the ASR transcription and the visual context.

To assess the performance of the multimodal system, another set of experiments will be conducted. In this case, the evaluation subset comprises previously unseen instances where speakers describe an image, and the described image will be extracted and used as input for the VLM to represent the user's visual context.

The objective of this experiment differs from previous objectives, as it shifts the focus from achieving accurate transcription to understanding the intended meaning when people with aphasia express specific sentences. As some types of aphasia present difficulties in finding the right words, limitations in speech, use of short sentences, or lack of coherence, in these cases, spoken sentences may not accurately convey intentions, introducing complexities in interpretation.

This renders the WER metric, previously used to compare transcription systems, no longer adequate, as a close resemblance to the reference sentence does not guarantee grammatical correctness or a full reflection of the user's intent. Therefore, to evaluate the effectiveness of this system, a human evaluation methodology has been adopted.

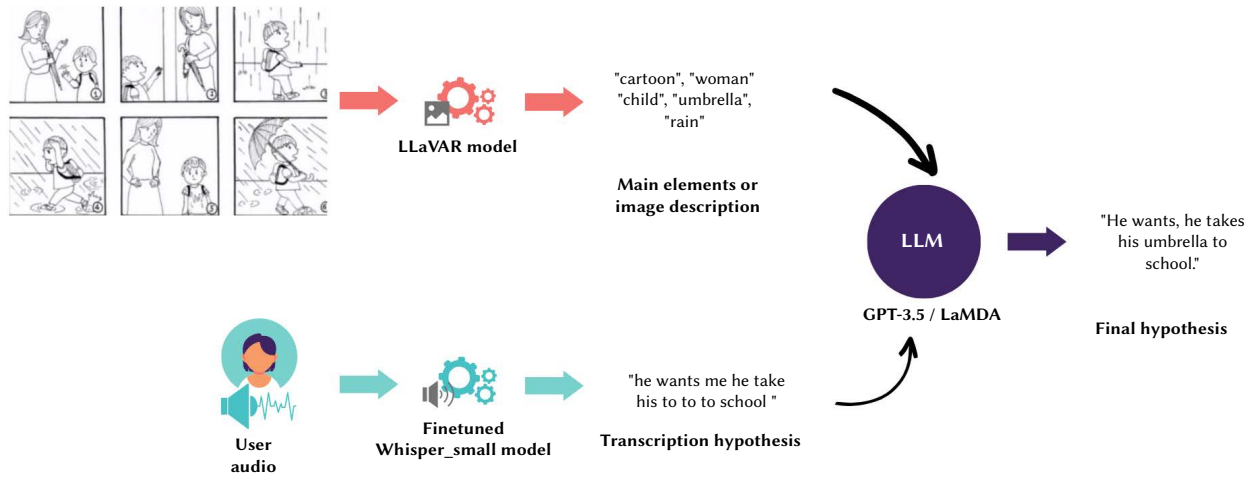


Fig. 3. Example of usage of the complete multimodal system given a specific image and audio.

For language model choice, GPT-3.5 [31] and LaMDA [32], two widely recognized models, have been tested. In the case of the image captioning system, LLaVAR [33] model has been selected for convenience and accessibility.

These models were selected as they represented, at the time of system development, some of the most advanced and widely accessible models available. Their strong performance in natural language understanding, combined with reliable public APIs, made them particularly suitable for integration into our multimodal pipeline.

Fig. 3 provides a specific example illustrating the process of the multimodal approach for one of the tested phrases in various experiments.

To conduct the human evaluation, a survey was administered to 80 individuals, with each question including audio spoken by a person with aphasia and transcriptions generated by some of the developed systems. Each of the presented audio segments had two questions in the survey. The first aimed to determine which transcription was more similar in terms of words to what the speaker expressed, while the second sought to identify which transcription better aligned with what the speaker intended to communicate.

Users provided feedback for each audio segment, resulting in a percentage score. The evaluation took into account the number of users who determined which model performed best for each specific question. Table V presents the survey results for the first question, where the most similar transcription to the audio was sought.

The results presented in Table V highlight that the system rated highest by surveyed users for providing the most similar transcription to the original audio is the Whisper fine-tuning, garnering 47.95% of user preferences.

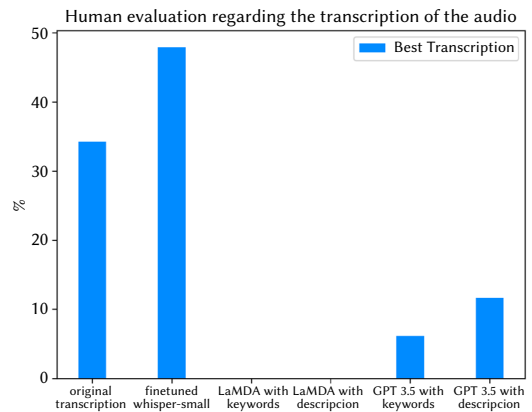
In contrast to the LaMDA-based system, those employing GPT as their language model received higher ratings from users.

In addressing the second inquiry, participants were asked to articulate their viewpoints regarding what the speaker tries to communicate in the sentences. Substantial discrepancies in results emerged when compared to the preceding experiment, notwithstanding the identical audio recordings and answer options. The detailed outcomes are illustrated in Table VI.

For this specific survey question, the same systems assessed in the prior experiment as possible choices have been incorporated. The inclusion of the original transcription of the audio from the dataset held particular significance in this instance, evaluating the extent to which a multimodal system could enhance its capacity to accurately

TABLE V. RESULTS OF HUMAN EVALUATION REGARDING THE SYSTEM THAT BEST REFLECTS WORD-FOR-WORD TRANSCRIPTION OF THE AUDIO

Model	Result(%)
original transcription	34.25
fine-tuned Whisper-small	47.95
LaMDA with keyword	0
LaMDA with description	0
GPT-3.5 with keyword	6.16
GPT-3.5 with description	11.64



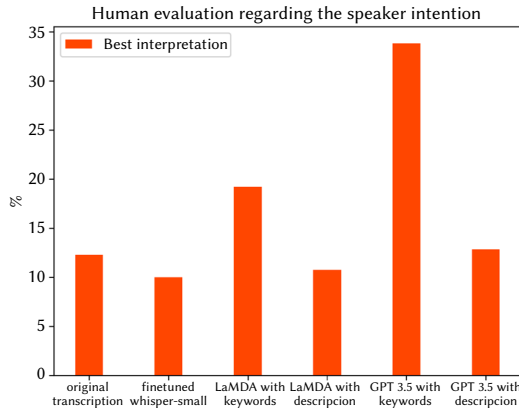
discern the user’s intended meaning compared to the reference phrase extracted from the dataset.

In this context, it is observed that the system yielding the highest performance, as indicated by 33.85% of respondents, is GPT-3.5 employed as a language model, supplemented with audio transcriptions and a list of elements present in the user’s visual environment. The second-highest rated system is the one utilizing LaMDA as a LLM, augmented with transcriptions and a list of primary elements in the accompanying image, garnering 19.23% support. The third position is occupied by the system employing GPT-3.5 as a language model, which receives visual information from the environment in the form of an image description, securing a 13.85% preference from users.

It is noteworthy that, in general, respondents opine that the transcriptions generated by these three models more accurately convey the intention of the speakers in comparison to the original audio transcription found in the dataset.

TABLE VI. RESULTS OF HUMAN EVALUATION CONCERNING THE SYSTEM THAT BEST REFLECTS WHAT THE SPEAKER INTENDS TO REFER TO

Model	Result(%)
Original transcription	12.30
fine-tuned Whisper-small	10
LaMDA with keyword	19.23
LaMDA with description	10.77
GPT-3.5 with keyword	33.85
GPT-3.5 with description	13.85



In this second experiment, a notable trend emerges, indicating that systems equipped with visual context information, presented through a list of key elements, consistently yield the most favorable results. It is worth mentioning that not only do systems utilizing 'keywords' outperform their counterparts, particularly those employing the same LLM with a descriptive approach, but systems provided with a list of keywords also receive more favorable ratings overall from users. This phenomenon can be attributed, among other factors, to the insight that presenting only the primary elements can offer crucial contextual information without introducing grammatical details that might confuse the system and alter the initial transcription.

VI. SYSTEM LIMITATIONS

After outlining the objectives, methodologies, and experiments, it is necessary to consider the limitations of this study. It is remarkable that the fine-tuned ASR model, one of the components of the presented multimodal models, has been trained with specific data from individual speakers exhibiting particular types of aphasia. Therefore, the efficacy of the model may be compromised when applied to users with different types of aphasia. Given the differences among speakers with aphasia and the considerable variability in their pronunciations, further adaptation may be necessary if a new speaker with distinct pronunciation impairments, not represented in the training data, were to utilize the model.

Although we are fully aware that dataset diversity and size are key factors in ensuring robust model generalisation, the availability of publicly accessible resources specifically focused on aphasic speech remains very limited. In this context, AphasiaBank represents a highly valuable and almost unique resource, as it provides real recordings of individuals with aphasia, along with time-aligned transcriptions, in both English and Spanish. Nevertheless, for future work, we consider it highly valuable to establish collaborations with individuals with aphasia or with organisations working with this community, with the aim of manually expanding the corpus and enriching it with new use cases in real-world settings.

Furthermore, the challenges posed by limited computational power, may have restricted the project's scope.

This work primarily focuses on the functional effectiveness of the system. However we are fully aware that, in order to facilitate its use in real-world assistive contexts, it is also essential to consider aspects such as computational cost, latency, and the feasibility of real-time deployment. The proposed system is based on the Whisper Small model, a relatively lightweight variant within the Whisper family, which has been specifically fine-tuned for recognising aphasic speech. This model can be executed efficiently on GPU-equipped servers or cloud instances without excessive resource demands. Regarding the visual modality, the system incorporates an image captioning model that can be deployed either on local GPU environments or via cloud-based services, depending on the resources available.

The total latency of the pipeline, comprising three sequential stages: ASR, image caption generation, and textual refinement via a language model, typically remains around 7 seconds when run on GPU-equipped environments (specifically, a NVIDIA GeForce RTX 4080), with an estimated maximum close to 10 seconds in cases of high processing load or particularly lengthy transcriptions. We consider this response time to be reasonable and acceptable for assistive applications, where the quality and reliability of communication are often more important than immediacy. Furthermore, inference optimisation, model quantisation, and the application of batch processing techniques could help to further reduce latency in future developments, and represent promising directions for ongoing and future research. Despite these constraints, the developments and experiments conducted have yielded remarkable results.

In summary, although the study is constrained by data availability and computational limitations, the proposed system has demonstrated promising results. It significantly improves automatic transcription of aphasic speech and shows the potential of integrating multimodal cues to enhance communication for individuals with expressive language impairments. These findings lay the groundwork for future refinements and larger-scale deployment.

VII. CONCLUSIONS

In conclusion, this project aimed to enhance user interaction with pronunciation issues, specifically focusing on aphasia in the context of ASR systems. The approach involved experiments incorporating fine-tuning of various models and the implementation of a multimodal system, combining fine-tuned ASR models with visual information and LLM for improved communication.

In the initial phase, a review of state-of-the-art projects and methodologies in ASR, image captioning, and LLMs was conducted, as well as the exploration and adaptation of the dataset used.

In the experiments, two approaches based on data partitioning were pursued, namely context-independent and speaker-independent. The context-independent approach compared several whisper models with Wav2Vec 2.0. Fine-tuning these models led to significant improvements, with Whisper-small achieving a 31.53% WER, representing a 55.19% relative improvement. The speaker-independent approach also demonstrated the effectiveness of fine-tuning, with Whisper-small achieving a 41.88% relative WER improvement.

Furthermore, the project addressed the goal of not just transcribing but effectively conveying the intention of messages from users with oral communication difficulties. Combining the best fine-tuned ASR system with visual information from a VLM alongside a LLM, such as GPT-3.5 and LaMDA, yielded promising results in human evaluations. Notably, 33.85% of users preferred the multimodal system using GPT-3.5 with a list of key elements for reflecting the speaker's intention.

It is worth noting that these findings could prove highly advantageous in the daily lives of individuals suffering from aphasia or facing difficulties in expressing themselves. Not only can they contribute to improving their interaction with technological devices, but they could also be instrumental in refining their communication with the broader society. In advanced cases of aphasia or other pronunciation disorders, it is common for affected individuals to encounter obstacles in communication. While those in their immediate surroundings who spend more time with them may, in some cases, better comprehend their intentions, understanding individuals with certain types of aphasia can be significantly challenging for those unaccustomed to interacting with such individuals.

Nevertheless, these systems, and research in this line of inquiry in general, could represent a substantial breakthrough for individuals with aphasia and oral expression difficulties. If the system can provide accurate transcription and effectively reflect what these users wish to express, it has the potential to overcome the communication barrier between these individuals and society.

In summary, the project successfully improved the performance of pre-trained ASR systems for individuals with aphasia, offering more accurate communication. The positive outcomes from human evaluations suggest the effectiveness of these systems. These findings provide optimism for future systems to leverage multimodal information, contributing to improved communication and inclusivity for individuals with disabilities like aphasia. The ongoing exploration of foundational models, fine-tuning, and model amalgamation remains crucial for enhancing interaction quality between individuals and technology.

CREDiT AUTHORSHIP CONTRIBUTION STATEMENT

Isabel Ferri-Molla: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Validation, Visualization, Project administration, Writing – original draft.

Jordi Linares-Pellicer: Conceptualization, Supervision, Project administration, Validation, Writing – review and editing.

Juan Izquierdo-Domenech: Validation, Supervision, Writing – review and editing.

DATA STATEMENT

The code and data processing scripts used in this study are publicly available at: <https://github.com/Vertexlit-VRain/aphasia>

The speech data used for training and evaluation were obtained from the AphasiaBank corpus (TalkBank). Due to ethical and privacy considerations, AphasiaBank data are not publicly available but can be accessed by researchers upon request through the official AphasiaBank data access procedure.

DECLARATION OF CONFLICTS OF INTEREST

No conflict of interest exists.

ACKNOWLEDGMENT

This work is partially supported by Generalitat Valenciana, FPI grant CIACIF/2022/098 and CI-PROM/2021/077.

REFERENCES

- [1] L. Rabiner, B. Juang, "An introduction to hidden markov models," *ieeE asSP magazine*, vol. 3, no. 1, pp. 4–16, 1986, doi: <https://doi.org/10.1109/MASSP.1986.1165342>
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017, doi: <https://doi.org/10.48550/arXiv.1706.03762>
- [3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023, pp. 28492–28518, PMLR.
- [4] A. Baeviski, Y. Zhou, A. Mohamed, M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020, doi: <https://doi.org/10.48550/arXiv.2006.11477>
- [5] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, *et al.*, "Seamless4tmassively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023, doi: <https://doi.org/10.48550/arXiv.2308.11596>
- [6] J. Li, *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022, doi: <https://doi.org/10.48550/arXiv.2111.01690>
- [7] M. Ozh, D. Oralbekova, K. Alimhan, M. Othman, B. Zhumazhanov, "Development online models for automatic speech recognition systems with a low data level," *Annals of Mathematics and Physics*, vol. 5, no. 2, pp. 107–111, 2022, doi: <https://doi.org/10.17352/amp.000049>
- [8] J. Nouza, L. Mateju, P. Cerva, J. Zdansky, "Developing state-of-the-art end-to-end asr for norwegian," in *International Conference on Text, Speech, and Dialogue*, 2023, pp. 200–213, Springer.
- [9] K. Deng, P. C. Woodland, "Adaptable end-to-end asr models using replaceable internal lms and residual softmax," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5, IEEE.
- [10] S. Dhahbi, N. Saleem, T. S. Gunawan, S. Bourouis, I. Ali, A. Trigui, A. D. Algarni, "Lightweight realtime recurrent models for speech enhancement and automatic speech recognition," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 6, pp. 74–85, 2024, doi: <https://doi.org/10.9781/ijimai.2024.04.003>
- [11] D. Mulfari, G. Meoni, M. Marini, L. Fanucci, "Machine learning assistive application for users with speech disorders," *Applied Soft Computing*, vol. 103, p. 107147, 2021, doi: <https://doi.org/10.1016/j.asoc.2021.107147>
- [12] N. Riccardi, S. Nelakuditi, D. B. den Ouden, C. Rorden, J. Fridriksson, R. H. Desai, "Discourse-and lesionbased aphasia quotient estimation using machine learning," *NeuroImage: Clinical*, vol. 42, p. 103602, 2024, doi: <https://doi.org/10.1016/j.nicl.2024.103602>
- [13] A. Adikari, N. Hernandez, D. Alahakoon, M. L. Rose, J. E. Pierce, "From concept to practice: a scoping review of the application of ai to aphasia diagnosis and management," *Disability and Rehabilitation*, vol. 46, no. 7, pp. 1288–1297, 2024, doi: <https://doi.org/10.1080/09638288.2023.2199463>
- [14] H. Yang, M. Zhang, S. Tao, M. Ma, Y. Qin, "Chinese asr and ner improvement based on whisper finetuning," in *2023 25th International Conference on Advanced Communication Technology (ICACT)*, 2023, pp. 213–217, IEEE.
- [15] J. R. Green, R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, *et al.*, "Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases," in *Interspeech*, 2021, pp. 4778–4782.
- [16] K. Rao, H. Sak, R. 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)*, 2017, pp. 193–199, IEEE.
- [17] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt, *et al.*, "Personalizing asr for dysarthric and accented speech with limited data," *arXiv preprint arXiv:1907.13511*, 2019, doi: <https://doi.org/10.48550/arXiv.1907.13511>
- [18] V. B. Kumar, S. Cheng, N. Peng, Y. Zhang, "Visual information matters for asr error correction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5, IEEE.
- [19] J. Effendi, A. Tjandra, S. Sakti, S. Nakamura, "Listening while speaking and visualizing: Improving asr through multimodal chain," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 471–478, IEEE.

- [20] S. Debnath, P. Roy, "Audio-visual automatic speech recognition using pzm, mfcc and statistical analysis," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, pp. 121–133, 2021, doi: <https://doi.org/10.9781/ijimai.2021.09.001>
- [21] S. K. Choe, Q. Lu, V. Raunak, Y. Xu, F. Metzke, "On leveraging visual modality for speech recognition error correction," 2019.
- [22] X. Chen, Y. Wang, X. Wu, D. Wang, Z. Wu, X. Liu, H. Meng, "Exploiting audio-visual features with pretrained av-hubert for multi-modal dysarthric speech reconstruction," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12341–12345, IEEE.
- [23] C. Yu, X. Su, Z. Qian, "Multi-stage audiovisual fusion for dysarthric speech recognition with pre-trained models," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1912–1921, 2023, doi: <https://doi.org/10.1109/TNSRE.2023.3262001>
- [24] E. Howarth, G. Vabulas, S. Connolly, D. Green, S. S. and, "Developing accessible speech technology with users with dysarthric speech," *Assistive Technology*, vol. 0, no. 0, pp. 1–8, 2024, doi: <https://doi.org/10.1080/10400435.2024.2328082>
- [25] G. Ayoka, G. Barbareschi, R. Cave, C. Holloway, "Enhancing communication equity: evaluation of an automated speech recognition application in ghana," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–16.
- [26] B. MacWhinney, "The talkbank project," *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, pp. 163–180, 2007, doi: <https://doi.org/10.1057/9780230223936>
- [27] B. MacWhinney, D. Fromm, M. Forbes, A. Holland, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011, doi: <https://doi.org/10.1080/02687038.2011.589893>
- [28] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, S. Hoi, "From images to textual prompts: Zero-shot visual question answering with frozen large language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10867–10877.
- [29] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, J. Luo, "Promptcap: Prompt-guided image captioning for vqa with gpt-3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2963–2975.
- [30] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplín, J. Heymann, M. Wiesner, N. Chen, *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018, doi: <https://doi.org/10.48550/arXiv.1804.00015>
- [31] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, *et al.*, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," *arXiv preprint arXiv:2303.10420*, 2023.
- [32] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022, doi: <https://doi.org/10.48550/arXiv.2201.08239>
- [33] Y. Zhang, R. Zhang, J. Gu, Y. Zhou, N. Lipka, D. Yang, T. Sun, "Llavar: Enhanced visual instruction tuning for text-rich image understanding," *arXiv preprint arXiv:2306.17107*, 2023, doi: <https://doi.org/10.48550/arXiv.2306.17107>



Juan Izquierdo Doménech

Juan Jesús Izquierdo Doménech is an Adjunct Professor of Computer Science in Universitat Politècnica de València (UPV, Spain). He received his Bachelor's degree in Computer Science Engineering from UPV and holds a Master's degree in Multimedia Applications from Universitat Oberta de Catalunya (UOC, Spain). He received his PhD in UPV in the field of Human-Computer Interaction, Mixed Reality, and Artificial Intelligence.



Isabel Ferri Mollá

Isabel Ferri Mollá is currently pursuing her PhD in Computer Science. She received her Bachelor's Degree in Computer Science Engineering from UPV and a Master's Degree in Artificial Intelligence, Pattern Recognition, and Digital Imaging (UPV, Spain). Her research interests include artificial intelligence, augmented reality, and human-computer interaction.



Jordi Linares-Pellicer

Jordi Linares Pellicer is an Associate Professor at Universitat Politècnica de València (UPV, Spain), where he leads the VertexLit research group at the Valencian Research Institute for Artificial Intelligence (VRAIN). He received his Ph.D. in Computer Science from UPV and holds a Master's degree in Artificial Intelligence from Universidad Internacional de La Rioja (UNIR, Spain).