racia

PILOT ACCEPTANCE EVALUATION RESULTS

Report

D3.3 – Pilot Acceptance Evaluation Results

Version Final – v2.1

November, 2024

AAL PROGRAMME

FCT Fundação para a Ciência e a Tecnologia

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Luxembourg National Research Fund

(Page intentionally blanked)

Project title	hOme-based Rehabilitation using an Artificial Companion for aphasIA
Project acronym	ORACIA
Grant agreement nº	AAL-2021-8-167-CP
Contract start date	1 March 2022
Contract duration	24 Months
Project coordinator	Instituto Pedro Nunes (IPN)

Document id (type)	D3.3 (REPORT)
Deliverable leader	RHZ
Due date	31/08/2024
Delivery date	31/08/2024
Dissemination level	Public (PU)
Status - version	Final – v2.1
Last update	22/11/2024

AUTHORS

Name	Organization
Margarida Realinho	IPN
Clara Szymanski	PSSJD
Marina Serena	PSSJD
David Benhsain	RHZ
Bianca Sousa	CRFT
Cristiana Fernandes	CRFT

PEER REVIEWERS

Name	Organization
João Quintas	IPN

REVISION HISTORY

Version	Date	Author/Organisation	Modifications
0.1	07.03.2024	Margarida Realinho / IPN	Creation from template
1.0	02.11.2024	David Benhsain / RHZ	Primary results
1.1	04.11.2024	Cristiana Fernandes / CRFT	Primary results
2.0	04.11.2024	João Quintas / IPN	Revision for FR
2.1	12.11.2024	Cristiana Fernandes / CRFT	Secondary results

Executive Summary

This document reports the results of the work executed in “Task 3.3 – Pilot Acceptance Evaluation Results”. This task run in parallel with the pilots with the objective of collecting the necessary information for the validation of the ORACIA implemented solution.

For the validation of the technology, we plan to collect some information about its use, make questioners to the end-users, caregivers, and other stakeholders to evaluate the social impact and the potential cost-effectiveness due to enhanced self-care, lifestyle, and care management.

ORACIA will be validated from the perspective of the co-design methodology, relating these activities to the discovery of relevant information about the usability of the solutions.

In terms of the pilots' usability assessment tools, we applied the instruments defined in D3.1. Additionally, this document also reports the main conclusions of pilot operation as defined in T3.2, which delivers D3.2 as a prototype.

Table of Contents

1. Introduction

With the ORACIA project, our primary goal is to create a product that uses technology to improve aphasia rehabilitation, thereby advancing the digital transformation of healthcare for people with aphasia. This effort primarily involves the development of novel software, designed and monitored by healthcare professionals, in conjunction with a digital application. The goal is to improve the process of aphasia rehabilitation.

This innovative solution, tailored for an elderly population, aims to foster collaboration between formal healthcare providers and informal caregivers by providing a unified Information and Communication Technology (ICT)-based platform. This platform will extend the reach of rehabilitation programs from clinical or hospital settings to community settings such as home care centers and patient residences.

The overall goal of ORACIA is to positively impact the healthcare sector by introducing a new service model. This model has the potential to reduce the overall healthcare costs associated with aphasia while alleviating the caregivers' burden in supporting individuals with this condition and allowing patients to have a more proactive approach in their rehabilitation.

During development, healthcare professionals tested the solution and accompanied new developments closely, reporting bugs and suggestions for improvement. In two different moments, 3 patients with aphasia also got to try ORACIA solution, providing early feedback and helping the development team understand the condition better to improve the interface and optimize usability.

The ORACIA project is expected to deploy three pilots in relevant environments (= total of 30 installations) that involved 115 end-users (45 primary end-users, 45 informal caregivers, 25 care professionals) to mainly validate the solution's user experience (UX), its feasibility and the adherence to the technology during the testing period of 6 weeks.

2. ASR system preliminary testing

The main component of the ASR system is the algorithm(s) that interpret the human speech utterances and infer the words it contains, known as Speech-to-Text (STT). This is a highly researched topic nowadays and every year new algorithms, solutions and applications are published or turned into consumer electronic products. However, the "best model" depends mostly on the circumstances, namely: language, speaker accent and pronunciation, hardware specifications, available processing resources, acceptable inference times, acoustics conditions, etc.

Besides the laborious task of gathering data, production of the ground truth (this annotation process is long), choosing/designing/tunning a customized model and training it, our option was to use already trained model. The approach was instead to test some publicly available STT models in our specific context and see which one(s) perform better. This document intends to illustrate the testing procedure and results.

The very first preliminary tests, uttering some of the chosen words, quickly revealed apparent reproducibility and selectivity issues, meaning that some STT models didn't produced the same result "output" when we repeatedly uttered the same "input". This didn't happen in all words, some got it consistently right and others wrong, but running two different sets of the same 27 different 'kitchen Tools' words uttered by the same speakers still showed this issue. Turning the climatization system of the room severely impacted some model's performance. Analysing two wav files of the same utterance just varying the silence portion duration on the edges (start and end) revealed relevant performance variation of the same model and surprisingly in the inference time of some models.

These issues demanded the need to be able to control the circumstances as much as possible in order to understand the impact of each factor – input conditions selectivity.

2.1. What is the "best model"?

In order to systematically test which speech-to-text (STT) model could better fit our purposes we created a pipe-line of python algorithms where we can test the models with different wav files. The approved list of words (divided in "items" like fruits, clothes, body parts, animals, tools, etc) were used to create wav files, first using readily available text-to speech (TTS) solutions, then using human test subjects and even testing different rooms.

These tests are design to assess several issues:

- Which one is the "best model"?
- Is it better for all cases?
- Is it fast enough or consistent?
- Is it the best for male and female speakers?
- Is the better model the best in all items (clothes, fruits, body parts, etc)?
- Is it the best for all languages?
- Are there some words (e.g., a specific fruit) that simply perform bad in all models?

- What impact has the rooms acoustics in the model's performance?
- Is the solution one simple "best model" or it depends on the selected language, gender of the speaker and can we build a "map" to the best model for each case?

Having the pre-requisite of only using off-line models, meaning the speaker utterance was only interpreted locally (and not sent to any online service), the current designed test routines were based on two "families" of STT models (which one with several variations):

- Vosk
- Whisper

2.2. Test Results

2.2.1. Using Synthetized Voices

The first approach to test the Speech-to-Test (STT) models without the influence of the microphone and rooms conditions, was to use speech utterances synthetized by algorithms of Test-to-Speech (TTS). The reproducibility of this approach can be confirmed by generating several times the same utterance, using the STT models and analysing the wav file the TTS algorithms. Not only can we inspect the wave file to be exactly the same if we generate it several times, but also confirm that the same model always infers (predicts) the same result. By listening to different test wav files containing utterances of fruits, clothes, body parts, etc., we can attest that the voice tone, intensity, pacing and other aspects are completely maintained across all wav files. Human speakers unintentionally produce variations on the referred aspects and that will produce performance variations that prevent us to test what is really the model performance.

A specific naming structure for each wav file was established along with a specific folder structure in order for the python scripts to run the tests in a systematic fashion, as shown in Figure 1.

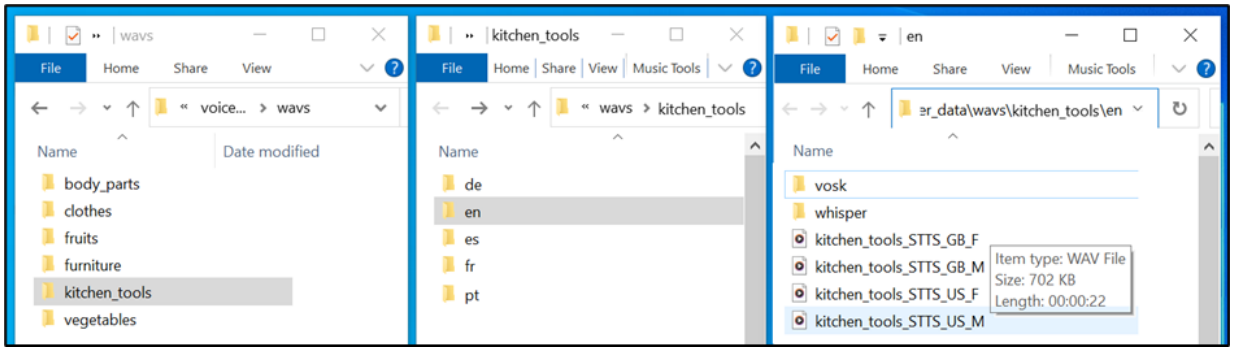


FIGURE 1. FILE STRUCTURE

The image exemplifies four files, each one containing the 27 approved words of the item 'kitchen_tools' uttered in English(en) in sequence (~22 seconds), where two are from a British speaker and the others from a United States speaker. The final character of the file names corresponds to the

gender of the speaker. Examining one of those on the time domain (Figure 2), we can identify (and listen) the uttered words and also the complete and evenly spaced silence between them (which in real conditions a human won't be able to reproduce).

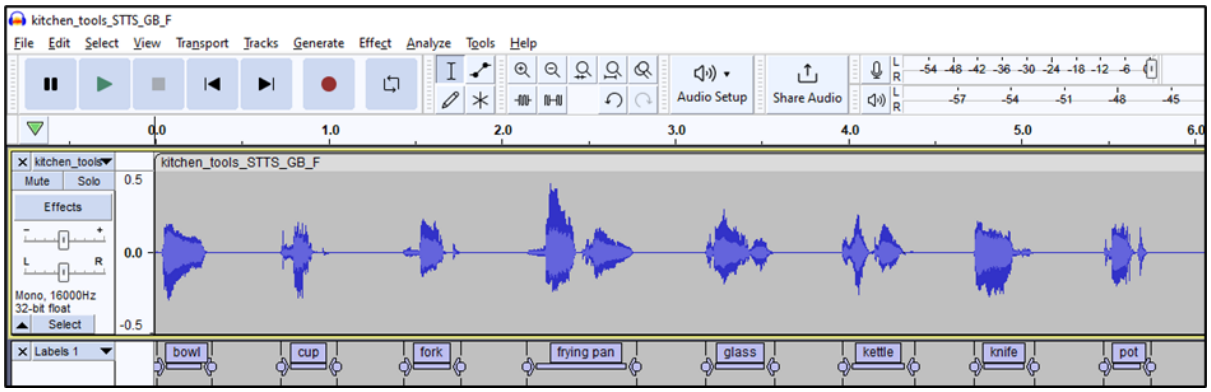
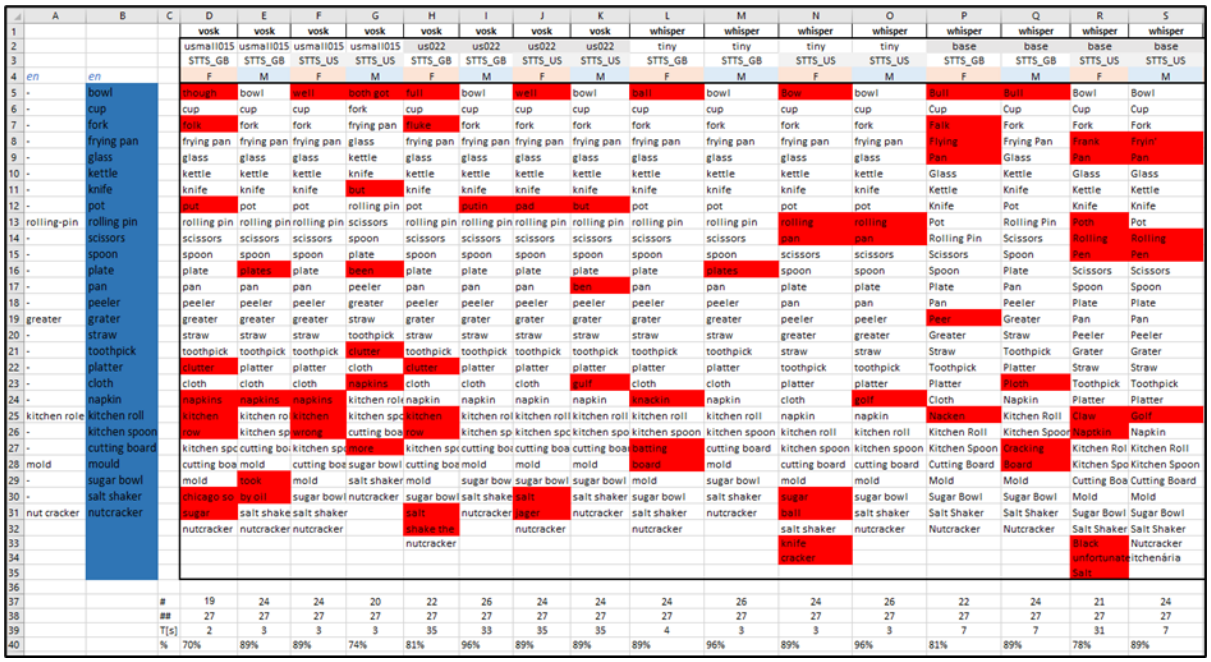


FIGURE 2. EXAMPLE OF WAVE FORMS FOR SYNTHESIZED UTTERANCES

The first test results, illustrated in Figure 3, allow us to understand which uttered word is correctly interpreted by each model but also to compare the percentage each different model inferred correctly (at the bottom), besides the wrong inferences (red background). Each four columns correspond to the same specific model runed on the 4 wav files (vosk_015, vosk_022, whisper_tiny, whisper_base and whisper_small).



			vosk	vosk	vosk	vosk	vosk	vosk	vosk	vosk	whisper	whisper	whisper	whisper	whisper	whisper	whisper	whisper
			usmall015	usmall015	usmall015	usmall015	us022	us022	us022	us022	tiny	tiny	tiny	tiny	base	base	base	base
			STTS_GB	STTS_GB	STTS_US	STTS_US	STTS_GB	STTS_GB	STTS_US	STTS_US	STTS_GB	STTS_GB	STTS_US	STTS_US	STTS_GB	STTS_GB	STTS_US	STTS_US
	en	en	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
5	-	bowl	though	bowl	well	both got	full	bowl	well	bowl	ball	bowl	Bow	bowl	Bull	Bull	Bowl	Bowl
6	-	cup	cup	cup	cup	fork	cup	cup	cup	cup	cup	cup	cup	cup	Cup	Cup	Cup	Cup
7	-	fork	folk	fork	fork	frying pan	fluke	fork	fork	fork	fork	fork	fork	fork	Falk	Fork	Fork	Fork
8	-	frying pan	frying pan	frying pan	frying pan	glass	frying pan	frying pan	frying pan	frying pan	frying pan	frying pan	frying pan	frying pan	Flying	Frying Pan	Frank	Fryin'
9	-	glass	glass	glass	glass	kettle	glass	glass	glass	glass	glass	glass	glass	glass	Pan	Glass	Pan	Pan
10	-	kettle	kettle	kettle	kettle	knife	kettle	kettle	kettle	kettle	kettle	kettle	kettle	kettle	Glass	Kettle	Glass	Glass
11	-	knife	knife	knife	knife	but	knife	knife	knife	knife	knife	knife	knife	knife	Kettle	Knife	Kettle	Kettle
12	-	pot	glut	pot	pot	rolling pin	pot	putin	pad	but	pot	pot	pot	pot	Knife	Pot	Knife	Knife
13	rolling-pin	rolling pin	rolling pin	rolling pin	rolling pin	scissors	rolling pin	rolling pin	rolling pin	rolling pin	rolling pin	rolling pin	rolling	rolling	Pot	Rolling Pin	Poth	Pot
14	-	scissors	scissors	scissors	scissors	spoon	scissors	scissors	scissors	scissors	scissors	scissors	pan	pan	Rolling Pin	Scissors	Rolling	Rolling
15	-	spoon	spoon	spoon	spoon	plate	spoon	spoon	spoon	spoon	spoon	spoon	scissors	scissors	Scissors	Spoon	Pen	Pen
16	-	plate	plate	plates	plate	been	plate	plate	plate	plate	plate	plates	spoon	spoon	Spoon	Plate	Scissors	Scissors
17	-	pan	pan	pan	pan	peeler	pan	pan	pan	ben	pan	pan	plate	plate	Plate	Pan	Spoon	Spoon
18	-	peeler	peeler	peeler	peeler	greater	peeler	peeler	peeler	peeler	peeler	peeler	pan	pan	Pan	Peeler	Plate	Plate
19	greater	grater	greater	greater	greater	straw	grater	grater	grater	grater	grater	greater	peeler	peeler	Peer	Greater	Pan	Pan
20	-	straw	straw	straw	straw	toothpick	straw	straw	straw	straw	straw	straw	greater	greater	Greater	Straw	Peeler	Peeler
21	-	toothpick	toothpick	toothpick	toothpick	clutter	toothpick	toothpick	toothpick	toothpick	toothpick	toothpick	straw	straw	Straw	Toothpick	Grater	Grater
22	-	platter	clutter	platter	platter	cloth	clutter	platter	platter	platter	platter	platter	toothpick	toothpick	Toothpick	Platter	Straw	Straw
23	-	cloth	cloth	cloth	cloth	napkins	cloth	cloth	cloth	gulf	cloth	cloth	platter	platter	Platter	Ploth	Toothpick	Toothpick
24	-	napkin	napkins	napkins	napkins	kitchen roll	napkin	napkin	napkin	napkin	knackin	napkin	cloth	golf	Cloth	Napkin	Platter	Platter
25	kitchen role	kitchen roll	kitchen	kitchen ro	kitchen	kitchen spc	kitchen	kitchen rol	kitchen roll	kitchen roll	kitchen roll	kitchen roll	napkin	napkin	Nacken	Kitchen Roll	Claw	Golf
26	-	kitchen spoon	row	kitchen sp	wrong	cutting boa	row	kitchen sp	kitchen spc	kitchen spo	kitchen spoon	kitchen spoon	kitchen roll	kitchen roll	Kitchen Roll	Kitchen Spoon	Naptkin	Napkin
27	-	cutting board	kitchen spc	cutting bo	kitchen sp	more	kitchen spc	cutting bo	cutting boa	cutting boa	batting	cutting board	kitchen spoon	kitchen spoon	Kitchen Spoon	Cracking	Kitchen Rol	Kitchen Roll
28	mold	mould	cutting boa	mold	cutting boa	sugar bowl	cutting boa	mold	mold	mold	board	mold	cutting board	cutting board	Cutting Board	Board	Kitchen Spo	Kitchen Spoon
29	-	sugar bowl	mold	book	mold	salt shaker	mold	sugar bow	sugar bowl	sugar bowl	mold	sugar bowl	mold	mold	Mold	Mold	Cutting Boa	Cutting Board
30	-	salt shaker	chicago so	by oil	sugar bowl	nutcracker	sugar bowl	salt shake	salt	salt shaker	sugar bowl	salt shaker	sugar	sugar bowl	Sugar Bowl	Sugar Bowl	Mold	Mold
31	nut cracker	nutcracker	sugar	salt shake	salt shaker		salt	nutcracker	jager	nutcracker	salt shaker	nutcracker	bail	salt shaker	Salt Shaker	Salt Shaker	Sugar Bowl	Sugar Bowl
32			nutcracker	nutcracker	nutcracker		shake the		nutcracker		nutcracker		salt shaker	nutcracker	Nutcracker	Nutcracker	Salt Shaker	Salt Shaker
33							nutcracker						knife				Black	Nutcracker
34													cracker				unfortunate	itchenária
35																	Salt	
36																		
37		#	19	24	24	20	22	26	24	24	24	26	24	26	22	24	21	24
38		##	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
39		T[s]	2	3	3	3	35	33	35	35	4	3	3	3	7	7	31	7
40		%	70%	89%	89%	74%	81%	96%	89%	89%	89%	96%	89%	96%	81%	89%	78%	89%

FIGURE 3. MODELS' PERFORMANCE ANALYSIS SUMMARY IN ENGLISH UTTERANCES

This particular test was repeated several times and revealed that vosk models tend to always give the same predictions and even took the same time (even the wrong ones were always the same). However, the whisper models (we only show the versions 'tiny' and 'base' but we also run the 'small' version) tend to perform differently if we run the same test over and over. Although it gets most of the words right, some are not always right and the ones it gets always wrong are not always the same wrong inference. This issue becomes more relevant because the inference times also can vary

considerably. The last 4 columns are the result of running the 'whisper_base' model on the 4 wave files but the second-last took much more time than the other three. The 'T[s]' row at the bottom corresponds to the time each model took the interpret the wave files with the 27 words utterances of kitchen tools and shows the vosk_022, which is a "bigger" model in relation to vosk_015, took 10x more time interpret the same wav files to interpret o average two more words correctly (out of 27). This cost relation, besides the need to be confirmed across more tests, can be relevant in a complete software interface where several processing demands can limit the available ones for this particular task.

Eventual performance patterns related to any preference of the models in relation to gender (M/F) or utterances made by British (GB) or United States (US) obviously need to be made in amore wide range of words as these particular 27 words may not be representative. Analysing the Portuguese language (Figure 4), which is a less common language than English and therefore has less available and less extensive STT trained models, the 4 wave files were subject to the same test structure. The difference is in the fact we used 2 synthetized utterances in Portuguese from Portugal and 2 others from Portuguese from Brazil (again, one male and other female). Other aspect is that the second vosk model (vosk_ptfb) is not just a "bigger version" of the first one (vosk_ptsmall), but a model trained on a larger dataset mostly from Brazilian utterances.

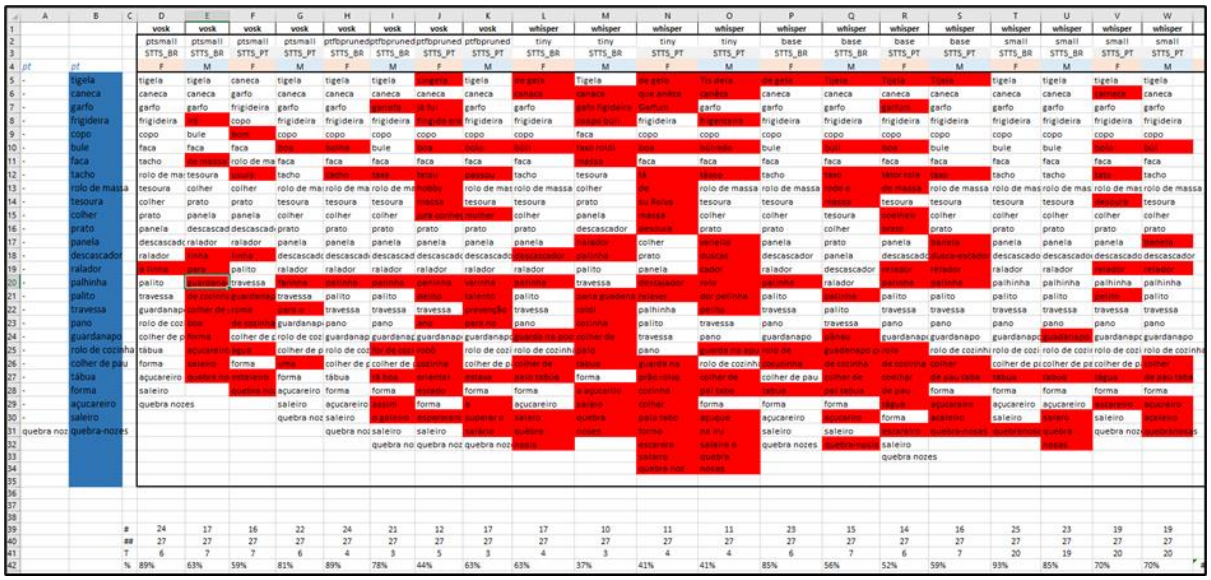


FIGURE 4. MODELS PERFORMANCE ANALYSIS SUMMARY FOR PORTUGUESE UTTERANCES

As it's visible the Portuguese version of the same tests produced more variations of the performance metric of the same families of STT models than in English (same architectures and similar sizes but trained in a different language). The vosk models seem to prefer the utterances that match the specific accent/dialect they were trained upon. The whisper models seem to also have that preference, and although their creators don't release details of their training dataset, it is reasonable to assume that they would more likely have access to Brazilian datasets.

2.2.2. Using human test subjects

For testing human voice utterance, when compared with the previous scenario, we have to add to the testing setup several "variables" all at once (a microphone, a speaker, a room, noise) of which we

would like to test the impact on the STT algorithms performance separately (or as separately as possible). Different microphone where tested, including the integrated in our laptop and even a headphone, but the more relevant is the one with the intrinsic capability of picking up the sounds from an entire room (an not just near ourselves). This introduced two aspects in the test scenario: the distance between the micro and the speakers and the room acoustics.

In this new stage of testing, because human speakers cannot control the speed they talk when reading a list of 27 kitchens tools in a row (producing very different durations of wav files to be interpreted by the STT models), the acquisition part of the testing process was done differently. Each word would generate one single wav file and, in this fashion, we are able to control and maintain a relatively constant "silence" in the beginning and end of the utterances (there cannot be abrupt cuts in a utterance and long silences led to wrong inferences). As referred the "silence parts" are no longer silence (as they were in the synthetized voice tests) and specially they are not constant nor predictable. These parameters changed the behaviour previously observed of the models and so new tests were made to identify the impact of each one.

The wave forms themselves enabled us to visualize not only that any human speakers is not capable of producing equal utterance over and over, but also that there is no longer true silence between utterances, as seen the following image Figure 5.

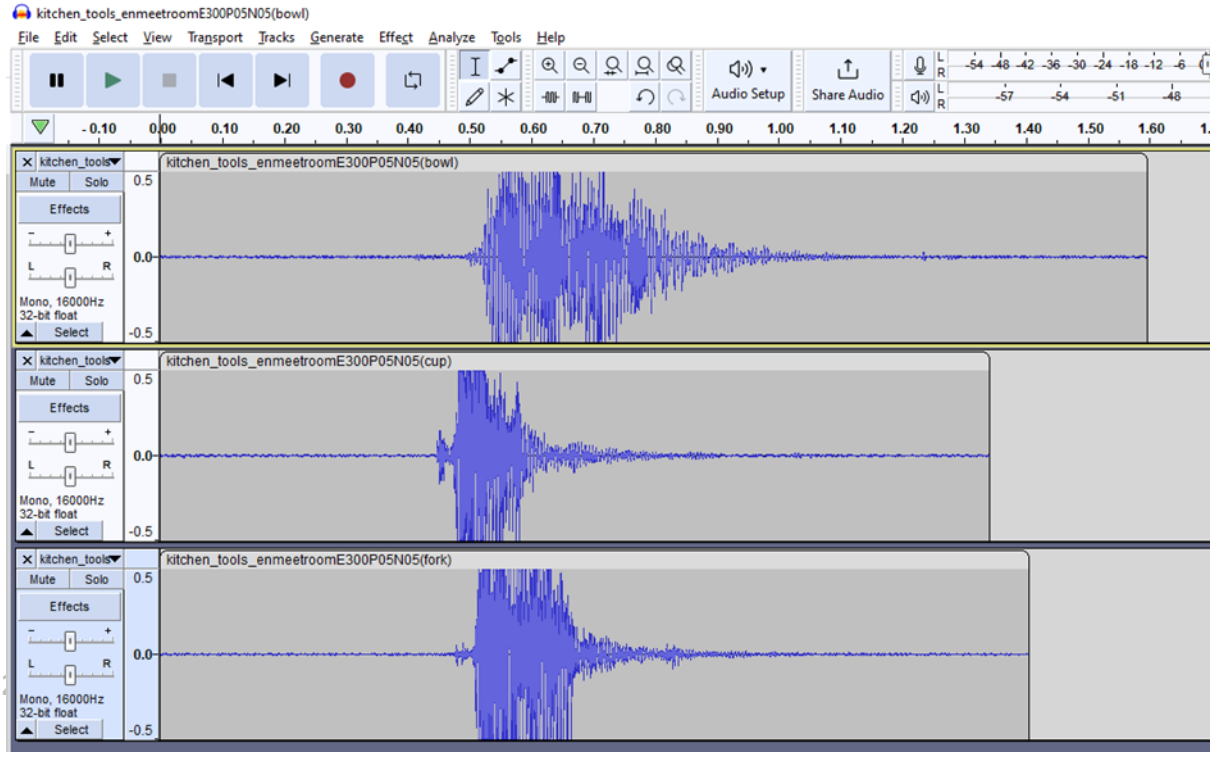


FIGURE 5. WAVEFORMS FOR HUMAN UTTERANCES

The strategy taken to produce these tests changed mostly taken in consideration the results themselves. The first approach of using synthetized voices happen because the first live human speech tests were producing very confusing results and the idea was to understand who was to blame – the acquisition process (microphone, room, software/libraries used, etc) or the models. This enabled to

understand what was the “best performance” possible if the conditions where “perfect” (which end up being the conditions most metrics obtained when the models where firstly published).

The most common metric used to compare the STT algorithms is Word Error Rate (WER), but in our specific application, where most uses of the STT models are for single word utterances, this metric offered little informative value. Using instead Character Error Rate (CER) proved to be a more effective and informative metric to asses the quality of the predictions and usefulness of each model (in each particular language). If the model “guessed” one long utterance except one or two characters we can still consider that result useful or at least infer that maybe it’s a pronunciation issue or background noise. As long as we perform extensive tests to check if that partial result is “stable” or repeatable, we can build a criteria (e.g. a simple threshold) to consider those as acceptable answers. This can be a valid approach as most scenarios we intend to use the ASR system there will be a very restrict context, meaning, we will know which is the correct answer and also important we can device cognitive exercises where the wrong answers are phonetically very different between themselves.

The fact that we understand some factors that influence each specific model to each specific language (and even factors like accent or gender) lead us to believe that the solution to this speech recognition task its not the “best model” but instead a “map of solutions” depending on the context. This map can be updated if we discover more STT models, or if we apply some “noise filtering” algorithm and make the wav files more “understandable” to the same algorithms. As is, the results indicate we are severely limited to use a very narrow list of each item type (fruits, clothes, body parts, etc), but the base is built to improve upon the current state of the ASR system performance.

The next sections also re-enforce the need to tailor the exercises to our specific context and limitation it produces. Homophone words need to be considered all the same “correct word” and so the exercise cannot have wrong options with those other homophone forms. If the task is to identify the ‘grater’ in a image we have to consider the “greater” inference returned by the STT model as a “correct answer”.

These tests raised many issues related to ambiguity where we may be forced to add new “acceptable answers” to one “expected word”. One odd linguistic situation related to these examples is the fact that some people, even when “faced” with all the reasons to identify “it” as the plural form, for instance an image of a bag ful of “feijões” (beans in Portuguese), many people still referred it as the singular form “feijão” (a single bean). For instance, “schuhablage” (shoe rack in German), has several variations: “schuhregal”, “Schuhbänke”, etc (with all their plural and other variation terms). All these terms are probably also understood by German speakers, even if one term is more consensual, and so we may need a list of all the possible vocables of each “correct word” of each item instance.

One aspect that became clear after the first STT models test results was that frequently there was more than one “correct inference” for each uttered word. This happens essentially because the languages used to test have homophone words. In our case, there’s the additional issue of having no context for any single simple utterance. Examples vary from language to language and there more common situations related to singular/plural forms (very common in French, e.g. ”tasses”/ ” tasse” or ”ciseau”/ ”ciseaux”) to extreme cases in English like “scissors” and “caesar's”. These examples and many others could only be identified after analysing these tests.

This leads to the concept that the ASR system based on the referred STT models will “decide” if the user uttered the correct word by looking into a list of “acceptable correct words” for each word on the lists of words. One example would be “joue” (cheek) whose plural would be “joues” that sounds

the same. However, “joues” sounds exactly as “jouer” which means “play” in English. So all these are to be accepted as correct if inferred by any of the STT models.

2.3.2. Hyphenation

A common semantic characteristic of many languages is the used of hyphen (-) to separate two words. This obviously raises the questions if of can we consider correct a inference that grammatically only differs from the presence (or not) of the hyphen. Apart from the fact that some cases both situations can be grammatically acceptable in a particular language, the principle of lack of context and the need to use a perfectly useful inference for our purposes push us to also consider the hyphen variations of a certain term to be all correct or acceptable answers. “Mini-skirt”/”miniskirt”(en), “mini-saia”/ “minisaia” (pt), “mini-jupe”/”mini-jupe”(fr) frequently are examples of this and others even can appear as two different words like “raincoat”/“rain coat”.

2.3.3. Other Ambiguities

Besides the referred ambiguities there’s other that were identified of which we added examples in the following table.

Clear Issues	Language	Acceptable examples	Not acceptable	Notes
Singular/plural	FR	“joue”/“joues” (cheek)	“jouer” (play)	All have the same pronunciation.
Hyphen-words	(several)	mini-skirt / mini skirt /miniskirt		All can be considered correct.
	FR	portemanteau/ portemanteaux porte-manteau/ porte-manteaux	(“porte manteau” can also be considered)	Same pronunciation
Special characters	FR	“cœur”/ “coeur”		Both are correct.
Dialect variations	PT(BR)	gabardine /Gabardina		Depends on training dataset or ground truth (Brazilian).
	PT(BR)	Alcachofra /alcaxofra		Has the same pronunciation but differs slightly in spelling.

TABLE 1. EXAMPLES OF WORDS AMBIGUITIES

Other cases exist where the decision to had them to the “acceptable words” may not be in favour. The next table has some examples.

Grey Issues	Lang.	Possible acceptable examples	Not acceptable examples	Notes
singular/plural	FR		sourcils /sourcils	Carefully pronounced and listen they are different, but ever so slightly.
	DE	(models mix them often)	Aubergine / auberginen	Just a the end (very common plural form in German).
Just differ in accentuation	PT		Pé, Pê	It differs on the vowel pronunciation.
Homophone words	EN	leek/leak	(Differs two character)	With no context we cannot know which one was “spoken”.
	EN	Thyme/time	(Differs one character)	Thyme vegetable has the same pronunciation of time.
	DE	(mushrooms /beer)	Pilz /pils	Same pronunciation.
Diacritic differences	ES		brócoli/ brocoli	Sometimes both are correct, but not in this case.
	ES		rábano / rabano	“Rabano” doesn’t exist!
	ES		calabacín / calabacin	“Calabacin” doesn’t exist!
Two words put together	ES	col morada/ colmorada	(Can we consider this as a criteria?)	They have the same exact sound…so can we accept them as “the same”?
	FR	Choufleur / chou-fleur	?	Many two-word terms come out together of the STT models!
	EN	Armchair /arm chair	?	Not the same but without context, as is our case, …
	DE	Kleiderständer /kleider ständer	(It doesn’t exist separatly but)	… it sounds the same and describes the object.
	DE	Massagetisch /massage tisch	(As in massage table)	It describes it, phonetically and etymologically

TABLE 2. EXAMPLES OF ALTERNATIVE ACCEPTABLE WORDS

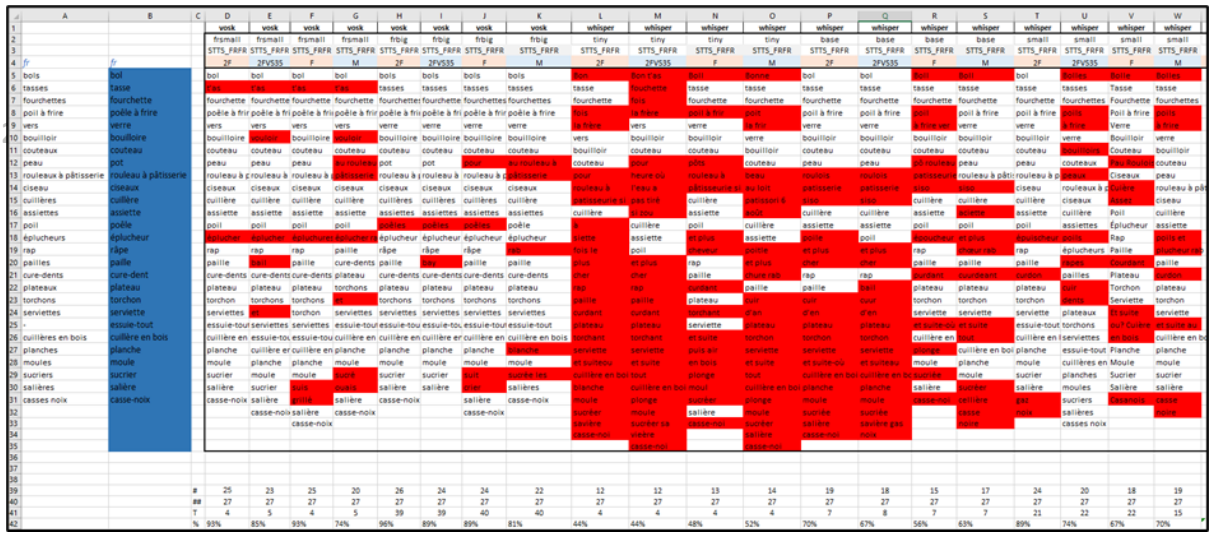


FIGURE 6. SYNTHETIZED UTTERANCES IN FRENCH

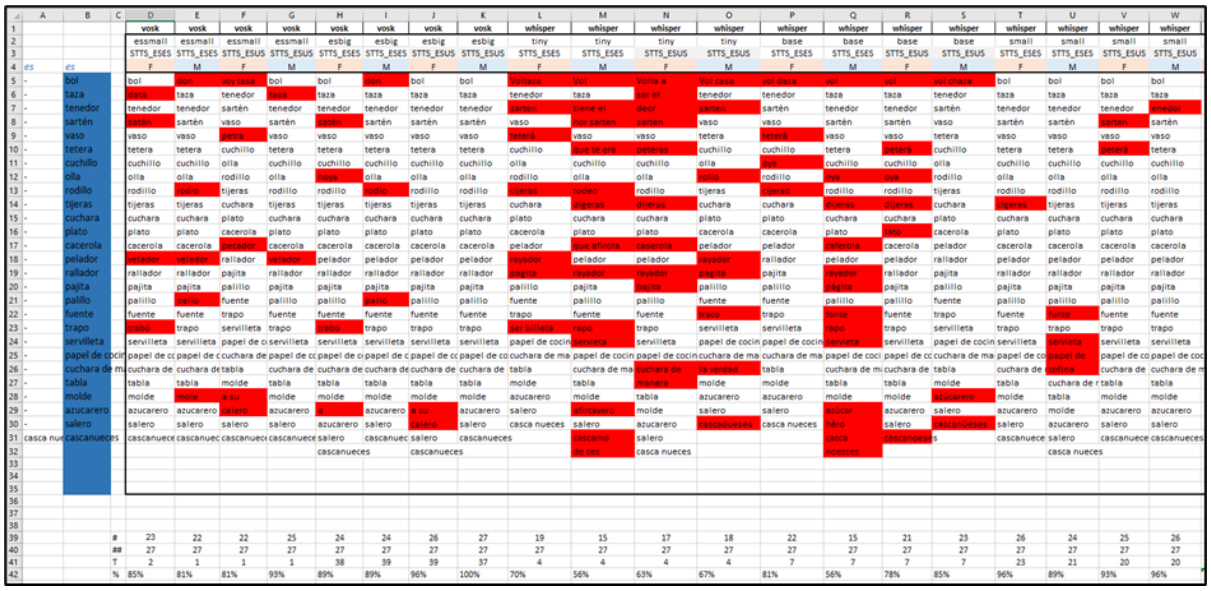


FIGURE 7. SYNTHETIZED UTTERANCES IN SPANISH

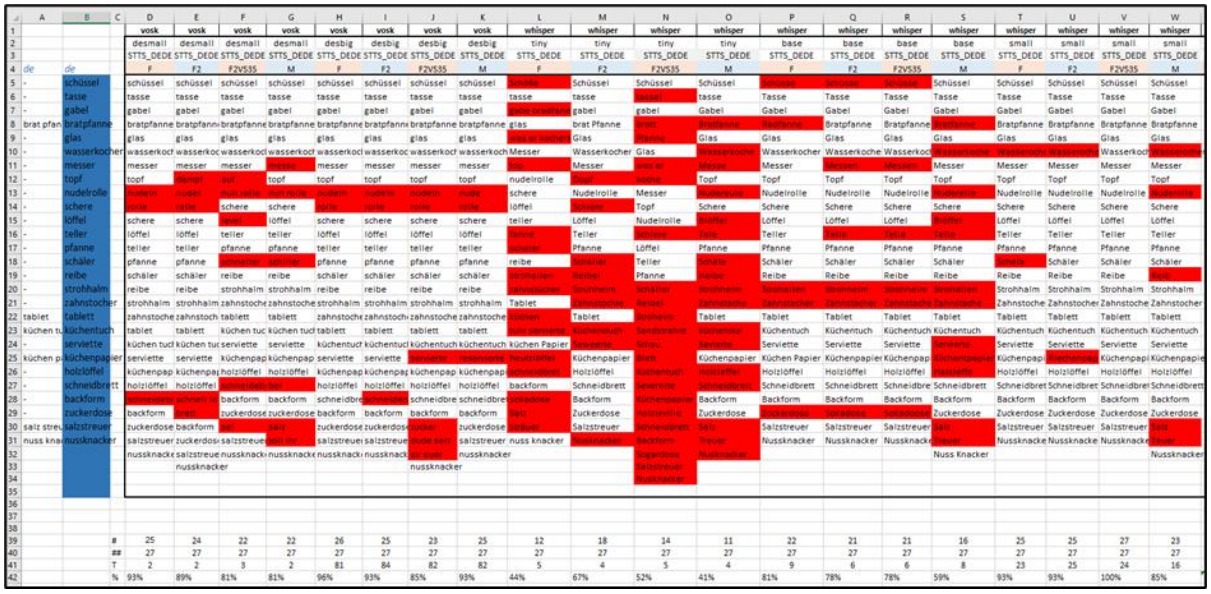


FIGURE 8. SYNTHETIZED UTTERANCES IN GERMAN

3. Overall system testing and debugging

The system was tested by partners along the development stage, every time a new feature or version was released. Bugs and suggestions for improvement were reported during weekly meetings and kept in the minutes. A bug/features status report was also kept in a common platform that can be accessed by the consortium members and technical research and development team. An example of the status of the bug report can be found in Annex 1.

A summary table for the expected involvement of participants in ORACIA usability study can as follows:

End-user types	CRFT	RHZ	PSSJD + Other	Total
Healthcare professionals	3 / 10	2 / 10	5 / 5	10 / 25
Patients	7 / 25	2 / 10	0 / 10	9 / 45
Total (HCP+Patients)	10 / 35	4 / 20	5 / 15	19 / 70
Caregivers	2 / 25	0 / 10	0 / 10	2 / 45
Total	12 / 60	4 / 30	5 / 25	21 / 115

3.1. Clínica de Recuperação Funcional da Trindade (CRFT)

Due to technical delays, the unavailability of the required number of prototypes to begin the study made it impossible to finish the pilot study before the final review. Nevertheless, given the necessity for ongoing improvement updates and considering that involving the final user from early stages can be extremely beneficial for the development of an adequate tool, two sessions were performed involving **3 therapists** and **3 patients** with aphasia, and **1 caregiver**, along with the research and development (R&D) team in clinic context. All sessions for testing ORACIA system follow the traditional methodology and last approximately 1 hour, always accompanied by a speech therapist.

In these sessions, the therapist accompanied the patients in a rehabilitation session using ORACIA, assisting in the navigation through the menus and trying different exercises. Feedback was collected by the R&D team and improvements were made in terms of usability and bug fixing, namely:

- Physical impairment on one side of the body can lead to difficulty using the touchpad on the keyboard. The interface required that the patient kept holding the microphone button while recording, releasing the button after speaking, and then submitting. This proved to be overly complex in terms of coordination for the patients and led to the alteration of the interface to allow for automatic recording;
- Some of the chosen images to represent the concepts were not adequate and led to confusion when naming the image shown;
- When the patient failed to give the correct answer, the platform used to give audio and visual feedback that the answer was wrong. The impact of this strong negative message proved to cause demotivation and anxiety to the player, adding to a fear of answering incorrectly. This wrong answer alert was then removed;
- A few bugs in the written information were detected and later fixed;
- The text-to-speech algorithm used for the interaction with the user was in Brazilian Portuguese, which posed a barrier to comprehension for patients who already struggled to understand the message. This algorithm was changed to accommodate European Portuguese;
- Overall, the speech-to-text algorithm failed to correctly detect the words spoken by the user, even when being spoken very clearly. This was a huge concern for the Portuguese language given that the model used is not as robust as in other languages. The solution was to introduce context to the words in the vocabulary to help the algorithm recognize them with more accuracy, which significantly improved the performance;

3.3.1 New results

- Involving an additional 4 patients (N=2 in clinic, N=2 in home) and 1 caregiver (i.e. linked to N=1 patient in home).
- Regarding images, there were some indications of improvement in certain concepts mainly in relation to quality, background, position and existence of sectioned images. The clinical team responsible for analyzing and selecting the images created by AI had already anticipated this

result, however, the development of the exercises and the consequent need to test the usability of the system implied the adoption of images of a different quality than the rest, which we hope to replace in the future;

- Despite exchanging some images and and indication of improvement of other, the rest of the vocabulary concept until now proved usabillity for comprehension and expression exercises, however clinical studies should continue and, in the future the sample size should be increased;
- The speech-to-text algorithm failed to correctly detect the words spoken by users with a larger number of paraphasias (mainly phonological) or with co-occurring motor control disorders such as apraxia or dysarthria (mainly flaccid). This increased difficulty in speech recognition was already anticipated, however we need to involve in the future a larger sample to collect more information and improve the system;
- Considering the communication difficulties of people with aphasia, it may be important to include in speech-to-text algorithm the word as the correct answer, whether it is plural or informal, e.g.: body parts or body part;
- Considering the use of ORACIA in a domestic context, it was noted that the surrounding noise is different from the clinical context, which may lead to different difficulties in speech-to-text algorithm. Despite the need to continue pilots to collect more information, testing using white noise such as a tumble dryer/washer, hair dryer, kitchen extractor fan, etc. should be considered;
- People with aphasia and caregivers valued health information, especially the guidelines and explanation of the concept of aphasia. While caregivers highlighted the importance of therapeutic guidelines for communicating, people with aphasia, especially younger ones and those with greater academic differentiation valued the accessibility of information in an aphasia friendly format. In general, they reported that although the guidelines had already been provided in a clinical context by the speech therapist, this written description and accessibility at any time in ORACIA were important. According to them, during treatment "there is so much information and bureaucracy", that it is "easy to forget what we heard" and it is "great" when I can "read the information at any time" or "share it with the family";

In general, despite the aforementioned bugs, people with aphasia were motivated and happy while using the "ORACIA" system. When asked if they would like to integrate the ORACIA system into the rehabilitation process, they responded affirmatively with: "of course yes"; "yes" and "can I use it now?". Users showed special motivation for using ORACIA at home, as according to them they normally use methods in paper format "exercises that the therapist gave" or videos in other languages "there were only Brazilian videos on YouTube", “I trained with what I had”.

From the caregiver perspective, the adoption of ORACIA system (tested only in web format) to support rehabilitation is very positive. According to her, she feels overwhelmed by all the information and sudden changes in her quality of life and routines, considering the support of the speech therapist and the possibility of carrying out the exercises that are prescribed to be very important, " without making the mistake of using what I shouldn't and make the situation worse."

In this specific case, the ORACIA system took on even greater importance because this is a case of a bilingual aphasic person (portuguese and english) who needs bilingual rehabilitation. Although we did not anticipate this positive result, it is motivating to realize from the speech therapist perspective that ORACIA system “supports the rehabilitation process, especially when the therapist is less fluency in one of the languages” and allows the realization of “monolingual therapy (where you focus on one language at a time) or bilingual therapy (where you work on both languages simultaneously or alternate between them)”. According to her, in bilingual therapy patients present "greater complexity in rehabilitation" as they require the “therapist have knowledge of both languages (speak and understand)”. Additionally, the “lack of adequate resources that integrate the necessary languages” represents a barrier to the rehabilitation process. With ORACIA she carry out “naming training using technology that integrates both languages and the same methodology, especially regarding the images and vocabulary used”. Another of the positive points highlighted is related to the possibility of continuing treatment even when the patient resides in another country.

FIGURE 9. PATIENT TESTING ORACIA IN CLINICAL CONTEXT.

FIGURE 10. SPEECH THERAPIST TESTING ORACIA IN HOME CONTEXT.

3.2. Rehazenter

Due to technical delays and the necessity for ongoing improvement updates, the pilot study could not be fully conducted before the final review. The system remains in a debugging phase, primarily tested by professionals to address the final adjustments required for its functionality. Despite these delays, the system was tested on **two patients and two professionals** at Rehazenter, yielding valuable insights into its current state and areas for improvement.

Observations from the testing indicate that the system has not yet reached a level of maturity for patients to use it independently. Currently, professional assistance is necessary, particularly for handling the microphone during naming exercises. While the system includes an extensive library of items, with a wide range of words and corresponding images in multiple languages (Portuguese, German, French, English, and Spanish), certain challenges remain. For example, aphasic patients often produce speech with "stuttering" or pauses, which the system currently interprets as errors. Although a patient may eventually produce the correct word, the system registers the initial "stuttering" period as a mistake, underscoring the need for further refinement in aphasic speech recognition. Additionally, the system is restrictive in recognizing answers; for example, it only accepts the isolated word "bottle" rather than a correct response in the form of a phrase, such as "it is a bottle."

Despite these technical challenges, the pilot yielded positive responses from both patients. They reported satisfaction with the interface and appreciated the clear and well-depicted items, expressing motivation to continue using the system for exercises at home. This positive feedback underscores the system's potential as an engaging tool for aphasic rehabilitation.

Feedback from professionals also highlighted areas for improvement. They noted the system’s potential but suggested refinements, particularly for the recording feature in naming tasks, to make it more intuitive and user-friendly. Additionally, the back-office interface, currently considered

rudimentary, could benefit from enhancements to facilitate the prescription of sessions and improve usability for healthcare providers.

In summary, while the ORACIA system demonstrates promising features and has garnered favorable feedback from patients and professionals, ongoing updates and refinements are essential to meet the unique needs of aphasic patients and streamline usability for independent home use.

At PSSJD, due to limitations stemming from a lack of national funding, the pilot with the ORACIA device is expected to involve only **5 healthcare professionals**. This pilot is scheduled for implementation at the end of 2024, as it is currently awaiting approval from the Ethics Committee. This testing phase will allow hospital care professionals to validate and provide feedback on ORACIA, as well as share their needs and insights based on their clinical experience. Their input will be invaluable not only for refining the device's functionality but also for tailoring it to fit seamlessly within real clinical workflows.

In preparation for the pilot, the project team conducted additional preliminary testing to identify and address any potential issues that project participants might encounter. This proactive testing ensured that participants could work with the latest, most stable version of ORACIA, allowing them to focus fully on the project's objectives without being hindered by avoidable technical issues or software bugs. The team's goal was to optimize ORACIA's reliability and ease of use, ensuring a smoother adoption during the pilot phase.

In this preliminary testing phase, the software version of ORACIA was utilized, with an emphasis on identifying areas for improvement. Key focus areas included:

- Adaptation for aphasic patients: Since the primary users of ORACIA will be patients with aphasia, it is crucial that voice recognition functions seamlessly, even under conditions where speech might be impaired. During testing, it was observed that voice recognition was not consistently effective, even with individuals who had no language impairments, raising concerns about its reliability for aphasic users. This limitation underscores the importance of fine-tuning the voice recognition algorithm to better accommodate varied speech patterns, phonetic deviations, and slower articulation. Ensuring accurate voice recognition is a vital component for the successful adoption and effectiveness of ORACIA, as it directly impacts patients' ease in using the device independently.
- Accessibility and usability: The device's instructions must be clear, concise, and contextually supportive, particularly given the language difficulties faced by aphasic patients. Preliminary testing revealed that many of the instructions on the ORACIA platform were either unclear or incomplete. For aphasic users, instructions need to be direct, unambiguous, and intuitively understandable to eliminate any doubts about how to use the device effectively. Addressing these usability gaps is essential for the device to be accessible, functional, and empowering for its intended users. Additionally, it was found that visual cues, such as icons and step-by-step prompts, could help bridge any gaps in understanding, making the device more intuitive for patients who struggle with text-based instructions.

- Feedback loop for continuous improvement: An iterative feedback mechanism was established to allow quick and continuous adjustments based on user feedback. The project team has committed to ongoing refinements, with updates planned in response to feedback collected from each testing phase. This dynamic approach aims to progressively enhance ORACIA's performance, particularly regarding accessibility and user experience, ensuring that it aligns with the evolving needs of aphasic patients and their caregivers.
- Clinical integration and training for professionals: Beyond usability for patients, the device's integration into clinical routines is crucial. Feedback from healthcare professionals indicated a need for comprehensive training to maximize ORACIA's potential benefits. By incorporating training sessions and user support resources, the project aims to ensure that clinicians can confidently guide patients in using ORACIA, fostering a collaborative and supportive environment that encourages patient autonomy.

These additional insights emphasize the commitment to creating a highly accessible, adaptable, and user-centered device that meets both the clinical needs of healthcare providers and the unique challenges faced by aphasic patients. Through this structured approach, the PSSJD pilot with ORACIA aims not only to validate the device's technical reliability but also to pave the way for broader, sustainable implementation within healthcare settings.

4. Conclusion

Although the pilots are not completed before the final review, the consortium has maintained an effort in testing and validating the solution since the beginning of the development process, involving therapists and patients when there was a possibility. At this stage, a total of 5 therapists and 5 patients were involved in different countries, and efforts will continue in order to complete the planned pilots, even after the end of this project. The consortium believes in the potential of ORACIA solution and aims to keep testing and developing a more robust platform.