Hochschule Niederrhein
University of Applied Sciences
iPattern
Institut für Mustererkennung
Institute for Pattern Recognition

Abschlussbericht
des
BMBF geförderten Vorhabens
TransLingo

Verbundprojekt: Transsektorale digitale Sprachtherapie (TransLingo)

Teilvorhaben an der Hochschule Niederrhein: *Dialoggestaltung mit robuster Spracherkennung für ein Aphasie-Sprachtrainingssystem*

Förderkennzeichen des BMBF: 13GW0481D

Laufzeit: 01.09.2021 bis 31.08.2025

Autor: Hans-Günter Hirsch

Datum: 10.11.2025

Vorbemerkung

Das Vorhaben trägt als Verbundprojekt die Bezeichnung „TransLingo“. Für das im Rahmen des Vorhabens entwickelte Sprachtrainingssystem wurde der Name „RehaLingo“ gewählt, um mit dem Kürzel „Reha“ für Rehabilitation die hauptsächliche Intention des entwickelten Systems zur Verbesserung und Wiederherstellung der sprachlichen Fähigkeiten bei Aphasie zu betonen.

Dieses Dokument wurde gemäß den Vorgaben des VDI als Projektträger der BMBF-Bekanntmachung ,,Psychische und neurologische Erkrankungen erkennen und behandeln - Potenziale der Medizintechnik für eine höhere Lebensqualität nutzen" als ausführliche Beschreibung der durchgeführten Arbeiten erstellt.

In dieser Darstellung der im Rahmen des Teilprojekts an der Hochschule Niederrhein durchgeführten Arbeiten und der dabei erzielten Ergebnisse wird auf die gendergerechte Schreibweise für Therapeut*in und Patient*in verzichtet. Wird im Text nur auf eine Patientin oder nur auf einen Patienten bzw. nur auf eine Therapeutin oder nur auf einen Therapeuten Bezug genommen, so sind diese Bezeichnungen in allen Fällen als geschlechtsneutral zu interpretieren.

Diese Dokumentation wurde ohne jegliche Zuhilfenahme von „large language models“ (LLM) erstellt.

Ausführliche Beschreibung der durchgeführten Arbeiten

Im Folgenden werden die im Rahmen des Vorhabens durchgeführten Arbeiten ausführlich dargestellt. Dabei werden zunächst die entwickelten graphischen Benutzerschnittstellen vorgestellt. Danach wird das Konzept und die Entwicklung der robusten Erkennung der sprachlichen Äußerungen von Aphasikern erläutert. Im Anschluss wird die konkrete Implementierung des Sprachtrainingssystems beschrieben. Zur Konfiguration der Erkennungssysteme und zum adaptiven Training wurde eine Sammlung von Aphasie Sprache erstellt. Der überwiegende Teil dieser Sprachdaten wurde bei Tests des Systems in einem Therapiezentrum aufgezeichnet. Die Ergebnisse einiger mit diesen Daten durchgeführten Erkennungsexperimente werden vorgestellt. Abschließend werden die Erfahrungen beim Test mit den Patienten beschrieben.

1. Graphische Benutzerschnittstellen (GUIs)
   Zunächst wird das grundsätzliche Konzept, das der Kommunikation des Sprachtrainingssystems mit dem Patienten zugrunde gelegt wird, vorgestellt. Es werden die grundsätzlichen Möglichkeiten einer Gestaltung des Sprachtrainings erläutert. Danach werden die GUIs zum Training des Wortverständnisses und der eigenständigen Sprachproduktion vorgestellt.

   1.1. Konzept der Dialoggestaltung
   Patienten führen ihr Sprachtraining in der Regel mit Therapeuten durch, die individuell auf die Eingaben des Patienten eingehen und ihm im Fehlerfall Hinweise auf die korrekte Eingabe geben. In diesem Teilprojekt wurde der Ansatz betrachtet, die Kommunikation zwischen dem Trainingssystem und dem Patienten dem Dialog zwischen Therapeuten und Patienten nachzuempfinden. Der Patient ist an diese Form der Therapie gewohnt, so dass die Hemmschwelle zur Akzeptanz des Systems gering ist und die Bedienung in der Regel kein großes Problem darstellt. Konkret werden kurze Videos der Therapeutin gezeigt, um einen Trainingsmodus vorzustellen oder die Reaktion der Therapeutin auf die Eingabe des Patienten zu präsentieren. Zudem sind Videos für etwa 200 Begriffe, die im Trainingssystem enthalten sind, vorhanden, in denen die Therapeutin den jeweiligen Begriff vorspricht. Aus den Bildern dieser Videos wurde der Bereich des Mundes extrahiert, um damit ein weiteres Video zu erzeugen, in dem die bewegte Lippenkontur bei der Aussprache des Begriffs zu sehen ist.

   Die Erstellung des Videomaterials war mit einigem Aufwand und dem Einbringen entsprechender personeller Ressourcen verbunden. Als Therapeutin wurde eine weibliche Person gewählt, die über die gesamte Laufzeit des Vorhabens zur Verfügung stand. In einem entsprechenden Aufnahme-Setup mit einem grünen Hintergrund („green screening“) wurden die Videos der Therapeutin erstellt, wobei dies nahezu über die gesamte Laufzeit des Projekts mit zunehmender Anzahl von Begriffen und Trainingsmodi in verschiedenen Aufnahmesitzungen erfolgte. Eingesetzt wurde eine Kamera des Typs „Sony ZV1“. Es wurden entsprechende Werkzeuge zur zeitlichen Beschneidung der Videos eingesetzt. Im Anschluss wurde mit entsprechenden Funktionen zur Bildbearbeitung der grüne Hintergrund entfernt. Des Weiteren wurden Funktionen der Matlab Bibliothek MTCNN [1] zur Bestimmung von 3 Orientierungspunkten in einem Gesicht (linker und rechter Mundwinkel sowie Nase) eingesetzt, um damit den Bereich des Munds über alle Bilder des Videos hinweg festzulegen. Damit wurde aus allen Bildern eines Videos der Bereich des Munds und damit der Lippenkontur extrahiert, um damit ein exakt gleich langes Video zu erzeugen, das die Bewegung der Lippen beinhaltet. Eine Herausforderung bestand noch darin, in einzelnen GUIs bei gleichzeitiger Wiedergabe des Therapeuten Videos und des Lippen Videos die beiden Videos zeitsynchron wiederzugeben. Dazu wurde in Matlab eine eigene Funktion erstellt, um die jeweils zugehörigen Bilder der beiden Videos gleichzeitig darzustellen.
   Die Bilder der etwa 200 Begriffe, die zum Training verwendet werden, wurden zum Teil durch das eigene Fotographieren von Gegenständen und zum Teil durch die Anwendung entsprechender KI-Werkzeuge zur künstlichen Generierung von Bildern erzeugt. Alle Bilder, auf denen Personen zu sehen sind, wurden künstlich generiert. Die zum Ende des Projekts hin verwendeten kurzen Videos, in denen eine Handlung dargestellt wird, wurden ebenfalls mit

entsprechenden KI-Werkzeugen künstlich erstellt.
Die etwa 200 Begriffe wurden für bestimmte Kategorien gewählt, denen sie zugeordnet werden können. Dabei wurden die nachstehenden Kategorien betrachtet, die Bereiche des alltäglichen Lebens abdecken:

- Küchenutensilien
- Wohnungsgegenstände
- Orte und Gegenstände im öffentlichen Raum
- Fahrzeuge
- Sportarten
- Tiere
- Pflanzen
- Gemüse
- Früchte
- Getränke
- Essensgerichte

Beim Start einer Trainingssitzung kann eine oder können mehrere Kategorien gewählt werden, aus denen die zu trainierenden Begriffe zufällig gewählt werden. Damit kann ein Schwierigkeitsgrad festgelegt werden. Werden beispielsweise mehrere gut zu unterscheidende Kategorien, z.B. Tiere und Fahrzeuge, gewählt, fällt die Unterscheidung von Begriffen aus unterschiedlichen Kategorien in der Regel leichter. Werden nur Begriffe einer Kategorie betrachtet, z.B. Küchenutensilien, ist die Aufgabe schwieriger. Ist ein Therapeut, der einen Patienten betreut, in die Gestaltung von Trainingssitzungen eingebunden, kann dieser aufgrund seiner Erfahrung mit dem Patienten auch gezielt bestimmte Kategorien auswählen. Zudem können Kategorien gemäß den persönlichen Interessen eines Patienten gewählt werden. Ein weiterer festzulegender Parameter, der die Schwierigkeit der Übung beeinflusst, ist die Anzahl der Begriffe, die in einer Trainingssitzung betrachtet werden.

Zum Training der eigenständigen Sprachproduktion, bei dem ein Patient einen Gegenstand benennt oder eine Szene sprachlich beschreibt, wird Spracherkennung eingesetzt. Ausgehend von dem Ansatz einer Reaktion des Systems, die der eines Therapeuten nahekommt, ist die Zielsetzung der Spracherkennung, den bei der Spracheingabe eines Aphasikers möglicherweise vorhandenen Artefakt zu bestimmen. Ein Therapeut reagiert auf die sprachliche Äußerung eines Patienten in Abhängigkeit dieses möglicherweise enthaltenen Artefakts. Dazu wird die nachstehende Zusammenstellung von Artefakten betrachtet, die in der entsprechenden Fachliteratur [2] beschrieben werden:

- Korrektes Wort ist enthalten neben Füllwörtern und Hesitationen
- Phonetisch ähnliches Wort
- Semantisch ähnliches Wort
- Oberbegriff, Überbegriff, Hypernym
- Unterbegriff, Hyponym
- Numerus, Singular, Plural
- Korrekte Eingabe des Worts mit signifikanter Pause zwischen Teilwörtern oder Silben
- Sehr lange Äußerung, in der auch das Zielwort enthalten sein kann
- Korrekte Wort ist nicht enthalten

Neben der korrekten Benennung eines Gegenstands oder Begriffs sind zwei typische Fehler eines Aphasikers die Äußerung eines phonetisch ähnlichen Worts oder eines semantisch ähnlichen Begriffs. Soll beispielsweise eine dargestellte „Maus“ benannt werden, könnte die Äußerung ein lautsprachlich ähnliches Wort wie „Laus“, „Haus“, „Moos“ oder „Maul“ beinhalten, wobei sich diese Beispiele jeweils nur in einem Laut im Vergleich zur korrekten phonetischen Beschreibung unterscheiden. Alternativ kann ein Patient einen semantisch ähnlichen Begriff äußern, z.B. „Ratte“ oder „Hamster“ im Fall des Zielworts „Maus“. Daneben passiert es auch, dass ein Oberbegriff, z.B. „kleines Tier“, oder ein Unterbegriff, z.B.

„Feldmaus“, genannt wird. Gelegentlich wird auch der falsche Numerus, „Mäuse“, geäußert. Bei zusammengesetzten Wörtern, z.B. „Radfahren“ als Bezeichnung einer Sportart, kommt es vor, dass die beiden Teilwörter mit einer signifikanten Pause dazwischen gesprochen werden. Bei bestimmten Formen der Aphasie kommt es zu sehr langen Spracheingaben.

Ein Patient äußert zur Benennung eines Gegenstands in der Regel nicht nur ein einzelnes Wort. Häufig sind Hesitationen oder Füllwörter enthalten, oder es wird eine komplette Phrase geäußert. Die Aufgabe der Spracherkennung ist es, in der Äußerung einen der vorgenannten Artefakte zu bestimmen. In Abhängigkeit des erkannten Artefakts reagiert das System mit einer dem Artefakt zugeordneten Reaktion des Therapeuten in Form eines kurzen Videos.

1.2. Trainingsmodi

Zum Wiedererlangen der sprachlichen Fähigkeiten und damit der Kommunikationsmöglichkeit mit Mitmenschen werden Begriffe und Gegenstände oder im späteren Verlauf der Aphasie-Therapie Szenen in drei Medienformen betrachtet, die in Abbildung 1 dargestellt sind.

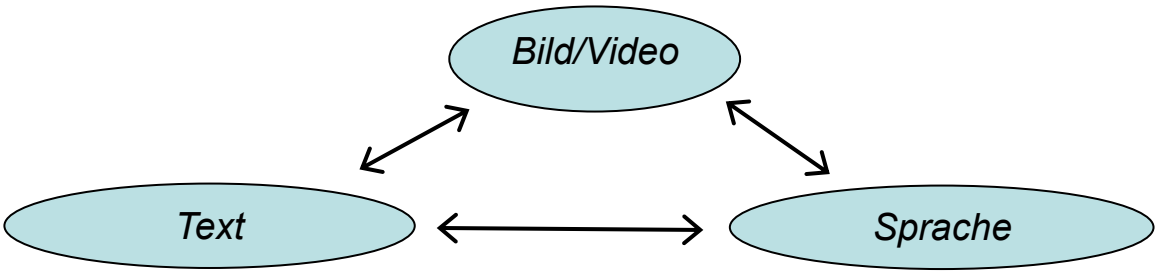

Abbildung 1: Medienformen zur Beschreibung eines Begriffs

Ausgehend von der Darstellung eines Begriffs in den drei Medienformen können eine Vielzahl von Trainingsformen und Trainingsformaten abgeleitet werden. Ein Begriff kann in einem oder in zwei Medienformen präsentiert werden mit der Zielsetzung, den Begriff in einer fehlenden Medienform zu beschreiben. In der herkömmlichen Therapie legt ein Therapeut beispielsweise ein gedrucktes Bild auf den Tisch und bittet den Patienten um die Benennung des dargestellten Gegenstands oder Begriffs. Zudem lassen sich eine Vielzahl von Varianten dieses grundsätzlichen Ansatzes ableiten. Beispielsweise kann ein Begriff in einer Medienform nur umschrieben oder nur zum Teil beschrieben werden, z.B. in Form eines Lückentextes, um aus dem Kontext auf den Begriff zu schließen. Die Aufgabe des Patienten ist die Beschreibung des Gegenstands in der gleichen oder in einer anderen Medienform. Ein Therapeut entscheidet in Abhängigkeit der speziellen Form der Aphasie und der individuellen Problematik bei einem Patienten, welche Trainingsformen in welcher Reihenfolge angewendet werden.

Im Rahmen dieses Vorhabens wurde nur eine Auswahl von Trainingsformen betrachtet. Zum Training des Verständnisses eines gehörten Worts oder Begriffs wurden graphische Oberflächen entwickelt, bei denen ein Begriff sprachlich über einen Lautsprecher wiedergegeben wird. Die Aufgabe des Patienten ist es, aus einer Auswahl von Bildern oder einer Auswahl geschriebener Wörter das zugehörige Bild oder den zugehörigen Text zuzuordnen. Diese beiden Trainingsformen wurden im Rahmen der Tests mit Patienten zur Einführung der in einer Sitzung zu lernenden Begriffe verwendet. Dabei hört der Patient die zu lernenden Begriffe in der Regel mehrfach, die er in den nachfolgenden Übungen eigenständig sprechen soll. Im Fokus des Projekts stand das eigenständige Sprechen eines Begriffs, da bei diesen Trainingsformen die Spracherkennung zur Analyse der sprachlichen Äußerung eingesetzt wird. Dabei handelt es sich für einen Patienten um eine wesentlich anspruchsvollere Aufgabe. Diese eigenständige sprachliche Beschreibung ist im Vergleich zum Wortverständnis wichtiger für die Verbesserung der Kommunikationsmöglichkeiten des Patienten. Im Folgenden werden die graphischen Oberflächen zum Wortverständnis und zur eigenständigen Sprachproduktion vorgestellt.

1.3. GUIs zum Wortverständnis

Die zum Training des Wortverständnisses entwickelten graphischen Oberflächen sind beispielhaft in den Abbildungen 2 bis 5 dargestellt. Zunächst wurden die GUIs unter Matlab entwickelt, die in den Abbildungen 2 und 3 dargestellt sind. Später wurde die Programmierung unter Python vorgenommen, da zum direkten Betrieb unter Matlab eine Lizenz erforderlich ist und die Einbindung der Erkennungsmodule unter Python mit geringerem Aufwand möglich ist. Die mit Python programmierten Oberflächen sind in den Abbildung 4 und 5 dargestellt, wobei jeweils ein Bild mit der Lippenkontur zu einem Zeitpunkt während der Wiedergabe des Therapeuten Videos zu sehen ist.
Bei allen Oberflächen zum Verständnistraining wird ein Begriff sprachlich über einen Lautsprecher wiedergegeben. Dazu wird in der Matlab Version zeitsynchron ein Video des Gesichts der Therapeutin und ein Video mit der extrahierten Lippenkontur wiedergegeben. In der Python Version wird nur das Video der Lippenkontur wiedergegeben, da sich herausstellte, dass die parallele Darstellung des Gesichts mit keinem sonderlichen Informationsgewinn verbunden ist. Der Patient kann sich das Wort durch Berühren des Lautsprecher Symbols oder eines der Video-Fenster erneut und beliebig oft anhören. Seine Aufgabe ist es, das dem wiedergegebenen Wort zugeordnete Bild oder den zugeordneten Text zu berühren. Nach der Auswahl und dem Berühren eines Bilds oder Texts erhält er eine akustische Rückmeldung in Form eines speziellen Tons für eine richtigen und eines anderen Tons für eine falsche Zuordnung. Zudem wird das Bild oder der Text mit dem richtigen Begriff mit einem grünen Rahmen und alle anderen Bilder oder Texte mit einem roten Rahmen versehen.
Während eines Trainingsdurchlaufs wird der Schwierigkeitsgrad adaptiv verändert. Dies erfolgt über die Anzahl der dargestellten Bilder oder Texte und ihrer Auswahl aus einer oder mehrerer Kategorien. In den Abbildungen 4 und 5 werden jeweils 5 Begriffe aus der gleichen Kategorie „Obst“ bzw. „Fahrzeuge“ angezeigt, da die Patientin zuvor den gesuchten Begriff bereits mindestens einmal richtig zugeordnet hatte. Bei der ersten Präsentation eines Begriffs oder nach einer falschen Zuordnung wird der Begriff mit einer kleineren Anzahl von alternativen Begriffen, die zudem aus anderen Kategorien stammen, präsentiert.
Der Fortschritt während eines Trainingsdurchlaufs wird über den orangen Balken im oberen Bereich der GUI veranschaulicht.

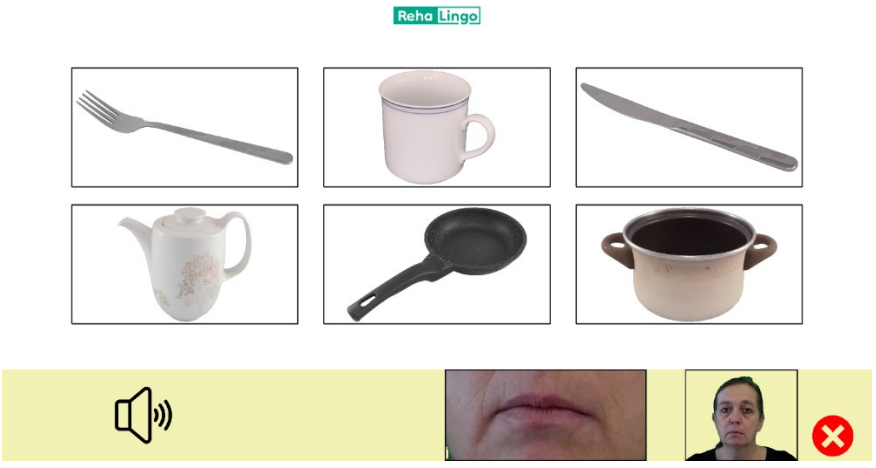

Abbildung 2: Graphische Oberfläche (Matlab) zum Wortverständnis mit Zuordnung eines Bilds

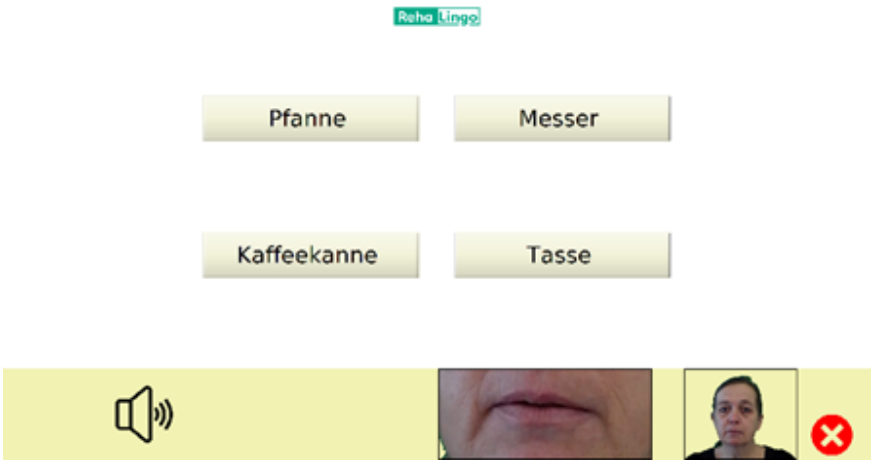

Abbildung 3: Graphische Oberfläche (Matlab) zum Wortverständnis mit Zuordnung eines Textes

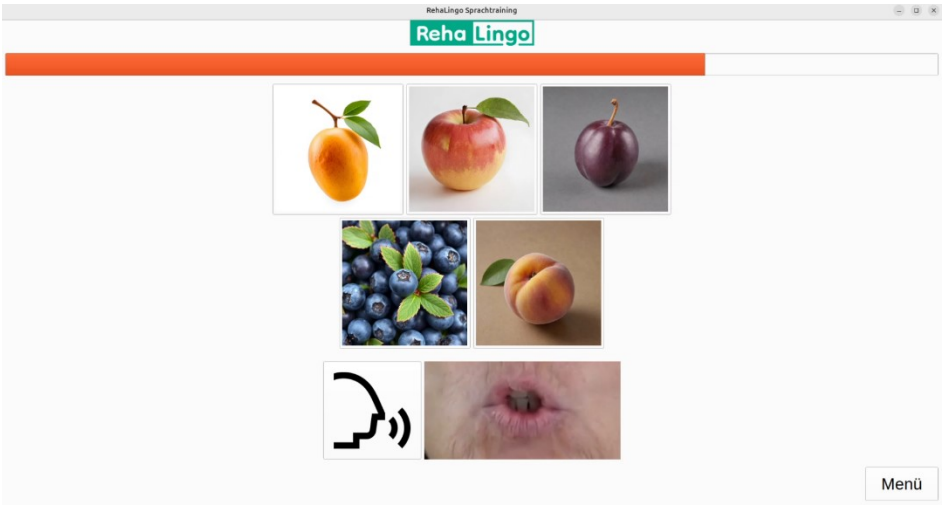

Abbildung 4: Graphische Oberfläche (Python) zum Wortverständnis mit Zuordnung eines Bilds

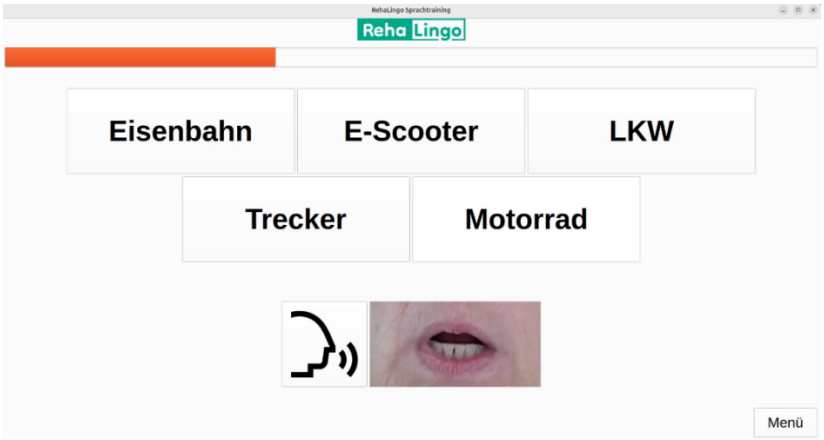

Abbildung 5: Graphische Oberfläche (Python) zum Wortverständnis mit Zuordnung eines Textes

1.4. GUIs zur Sprachproduktion

Es wurden mehrere GUIs zum Training des Sprechens von Begriffen entwickelt. Sie unterscheiden sich in der Vorgabe des zu sprechenden Begriffs in verschiedenen Medienformen. Einige Beispiele der GUIs sind in den Abbildungen 6 bis 8 dargestellt.
Bei der Version, bei der die meisten Informationen in Form eines Bilds, des zugehörigen Texts und der Wiedergabe des Worts über Lautsprecher vorgegeben werden, handelt es sich um eine einfache „Nachsprech“übung. Der Begriff wird in allen drei Medienformen präsentiert und soll anschließend sprachlich geäußert werden. Daneben gibt es mehrere Versionen, bei denen der Begriff in Form eines Bilds und eines vollständigen oder unvollständigen Texts präsentiert wird. In Abbildung 6 ist eine Version dargestellt, bei der neben dem Bild eine Schriftkontur mit einigen darin enthaltenen Buchstaben vorgegeben wird. Die Kontur gibt vor, mit welcher Höhe die fehlenden Buchstaben normalerweise geschrieben werden. Damit wird das äußere Bild der Textdarstellung als weitere Assoziationshilfe vorgegeben. In den Abbildungen 7 und 8 sind zwei Beispiele dargestellt, bei denen der gesuchte Begriff nur in Form eines Bilds präsentiert wird. Abbildung 7 beinhaltet ein Bild während der Wiedergabe des Videos, in dem die Therapeutin die korrekte Benennung des Begriffs erläutert. In dem Fall der korrekten Beantwortung wird das Bild mit einem grünen Rahmen versehen. Abbildung 8 beinhaltet ein Bild während der Wiedergabe des Videos, in dem die Therapeutin den detektierten Aphasie Artefakt erläutert, der in diesem Beispiel in der Äußerung eines Oberbegriffs bestand („es gibt noch eine genauere Bezeichnung“). Das Bild wird einem orangen Rahmen versehen, um die nicht ganz korrekte Benennung zu visualisieren. Im Fall einer falschen Eingabe wird das Bild mit einem roten Rahmen dargestellt.

Abbildung 6: Graphische Oberfläche zur Sprachproduktion bei Präsentation eines Bilds und eines Lückentextes mit einer Schriftkontur

Zur sprachlichen Eingabe berührt der Patient die Fläche mit der Beschriftung „Sprechen“ und dem Mikrofonsymbol. Dann wird das Signal des Mikrofons über einen festzulegenden, maximalen Zeitraum aufgezeichnet. Bei der erwarteten Eingabe eines einzelnen Worts wird dabei ein Wert im Bereich von 10 Sekunden gewählt. Des Weiteren wird eine Funktion zur automatischen Detektion des Endes einer sprachlichen Eingabe aktiviert. Man bezeichnet dies als „voice activity detection“ (VAD). Erkennt der VAD eine Sprachpause mit einer festzulegenden zeitlichen Länge, wird die Aufnahme vorzeitig beendet, so dass die Auswertung mit den Spracherkennungsmodulen stattfinden kann. Die Pausenlänge wurde für die Eingabe eines Begriffs zu 1 s gewählt. Die maximale Länge einer Aufnahme und die minimale Länge der Sprachpause für den VAD sind bei der Spracheingabe von Aphasikern kritische Parameter. Patienten äußern häufig nicht nur das gesuchte Zielwort und machen während einer Eingabe möglicherweise eine kurze „Denkpause“. Daher ist es sinnvoll, für beide Parameter größere Werte im Vergleich zu einer „normalen“ sprachlichen Eingabe zu

wählen. Eine Erhöhung der zu detektierenden Pausenlänge hat allerdings die Konsequenz, dass die Ergebnisse der Erkennung erst mit einer entsprechenden zeitlichen Verzögerung zur Verfügung stehen. Damit erfolgt auch die Reaktion des Systems in Form eines Videos der Therapeutin verzögert. Ein Mensch toleriert nur Verzögerungen bis zu einer gewissen Länge. Daher muss für die Festlegung der zeitlichen Länge zur Erkennung des Endes einer Spracheingabe ein Kompromiss gefunden werden, bei dem die sprachliche Eingabe nicht zu oft beschnitten wird und die Reaktion noch mit einer tolerablen Verzögerung erfolgt. Eine weitere Möglichkeit besteht in der Festlegung des Werts als wählbaren Konfigurationsparameter einer Trainingssitzung, so dass er individuell für jeden Patienten angepasst werden kann.

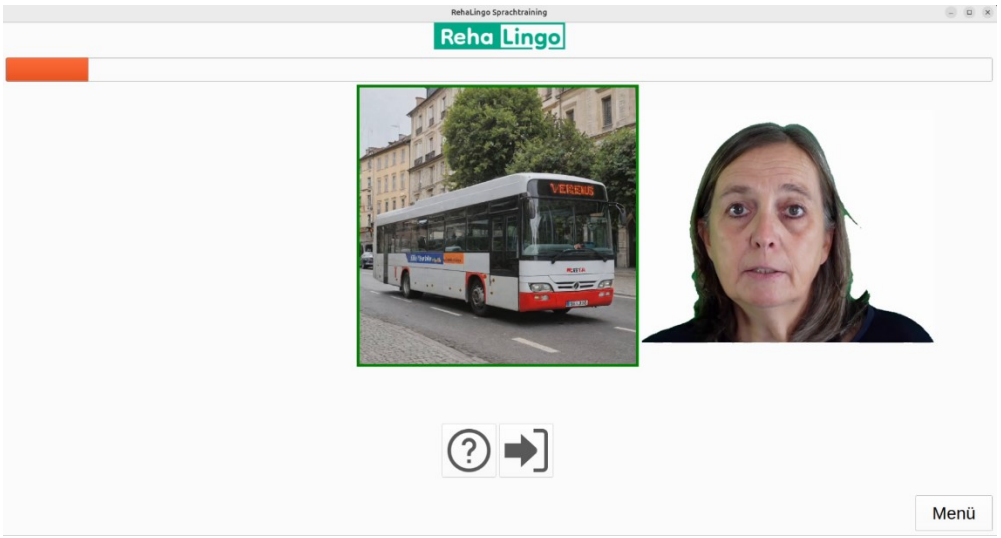

Abbildung 7: Graphische Oberfläche zur Sprachproduktion bei Präsentation eines Bilds

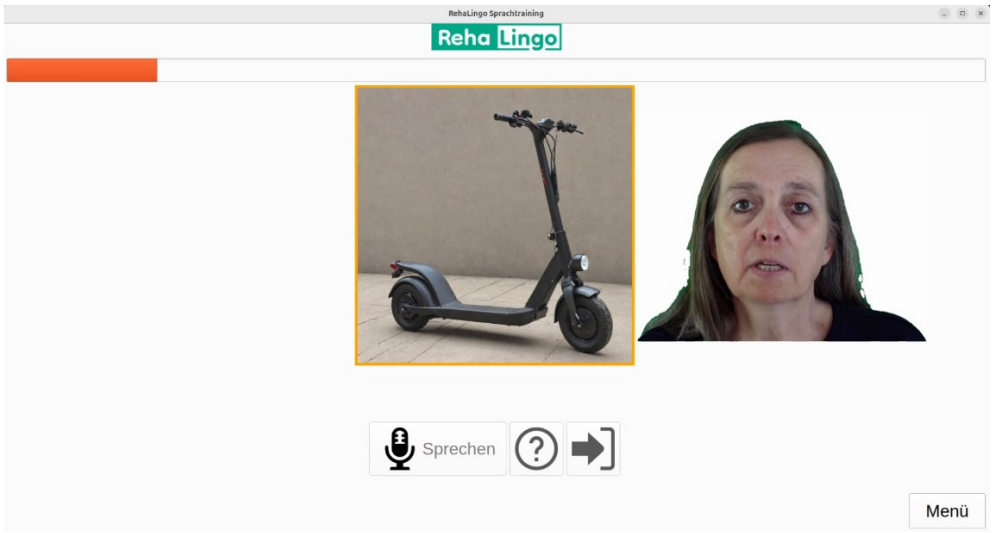

Abbildung 8: Graphische Oberfläche zur Sprachproduktion bei Präsentation eines Bilds

Zur Kommunikation der GUI mit den Spracherkennungsmodulen wurde eine softwaremäßige Schnittstelle in Form einer Speech-API entwickelt. In der Speech-API werden die Ergebnisse der Erkennungsmodule analysiert. Bei einer nicht korrekten Eingabe wird die Zugehörigkeit zu einem der in 2.1.1. vorgestellten Aphasie Artefakte vorgenommen. Die Speech-API liefert der GUI neben allen Erkennungsergebnissen in Textform die Information über den detektierten Aphasie Artefakt. In der GUI wird dann ein dem jeweiligen Aphasie Artefakt zugeordnetes Video der Therapeutin abgespielt. Der Patient hat in diesem Fall die Möglichkeit einer erneuten Spracheingabe, indem er nochmals die „Sprechen“ Fläche berührt.

Neben den GUIs zum Training von Begriffen wurde im Jahr 2025 eine GUI entwickelt, in der kein Bild, sondern ein kurzes Video einer Szene gezeigt wird. Ziel ist die sprachliche Beschreibung der Szene. In dem in Abbildung 9 dargestellten Beispiel wird ein etwa 4s langes Video gezeigt, in dem ein Mann mit einem Hund auf einer Straße geht.

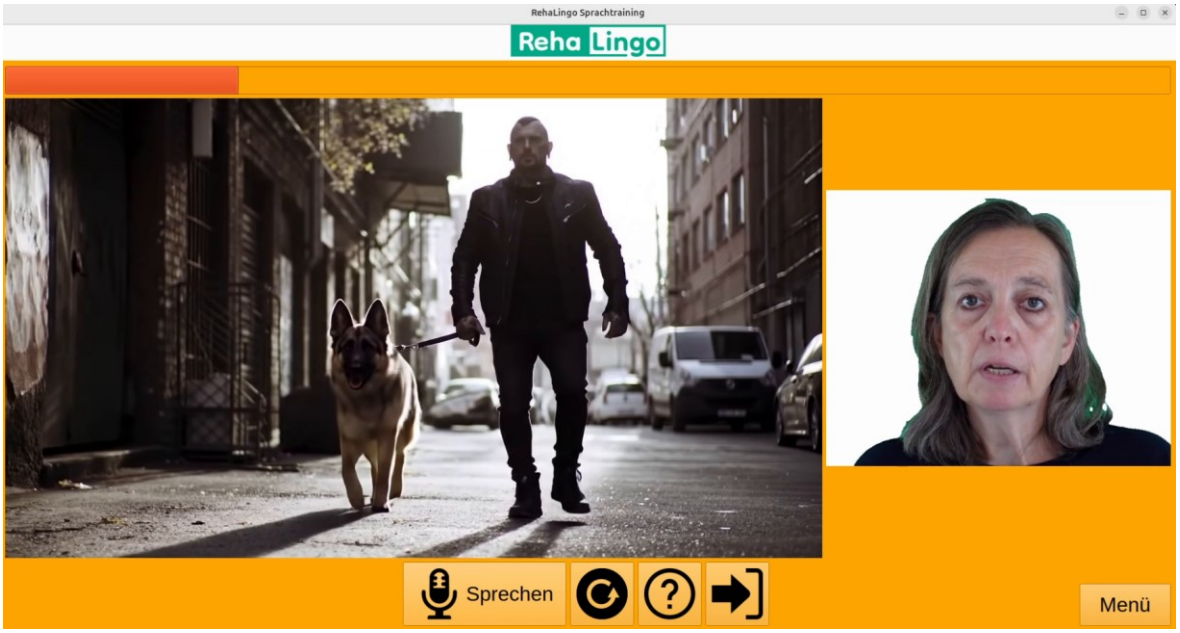

Abbildung 9: GUI zur sprachlichen Beschreibung einer kurzen Szene

Diese Aufgabe ist für die meisten Aphasiker deutlich schwieriger im Vergleich zur Benennung eines Gegenstands. Die sprachliche Beschreibung dieser Szene besteht häufig nur aus der Aneinanderreihung einiger Wörter wie „Mann Hund gehen“. In diesem Trainingsmodus wird nur das Whisper Modul zur Spracherkennung benutzt. Die von Whisper erkannte Wortfolge wird mit einem zusätzlichen Modul analysiert, das ein „natural language processing“ (NLP) beinhaltet. Das NLP Modul wurde so konfiguriert und trainiert, dass die Handlung als „Intent“ und die handelnde Person, ein Mann oder eine Frau, als „Entity“ bestimmt werden. Bei einer beispielhaften Erkennung der Wortfolge „Da geht jemand mit einem Hund“ wird die Handlung richtig beschrieben, also der „intent“ wird erkannt. Die „entity“ und damit die Aussage, dass es sich bei dem jemand um einen Mann handelt, wird nicht richtig benannt. In Abbildung 9 ist ein Bild dargestellt, das zu dem Video gehört, in dem die Therapeutin als Reaktion auf die zuvor genannte Äußerung die richtige Beschreibung der Handlung bestätigt und zusätzlich einen Hinweis gibt, dass der Patient die handelnde Person genauer beschreiben soll. Als Hintergrundfarbe wird in diesem Fall „orange“ gewählt, um die richtige Eingabe des „Intent“, aber die noch nicht korrekte Beschreibung der „Entity“ zu visualisieren. Bei einer korrekten Eingabe von „intent“ und „entity“ wird grün, bei einer falschen Eingabe rot als Hintergrundfarbe gewählt. Der Patient hat an dieser Stelle mehrere Möglichkeiten, wie das Training und der Dialog fortgeführt werden:

- Erneute sprachliche Eingabe
- Erneute Wiedergabe des Videos
- Video der Therapeutin mit korrekter Beschreibung der Szene
- Fortführung des Trainings mit dem Video einer anderen Szene

1.5. GUI zur Konfiguration eines Trainingsdurchlaufs

Um bei Tests mit Patienten einen Trainingsdurchlauf zu konfigurieren, wurde die in Abbildung 10 dargestellte graphische Benutzerschnittstelle erstellt, die von der eine Trainingssitzung begleitenden Person bedient wird.

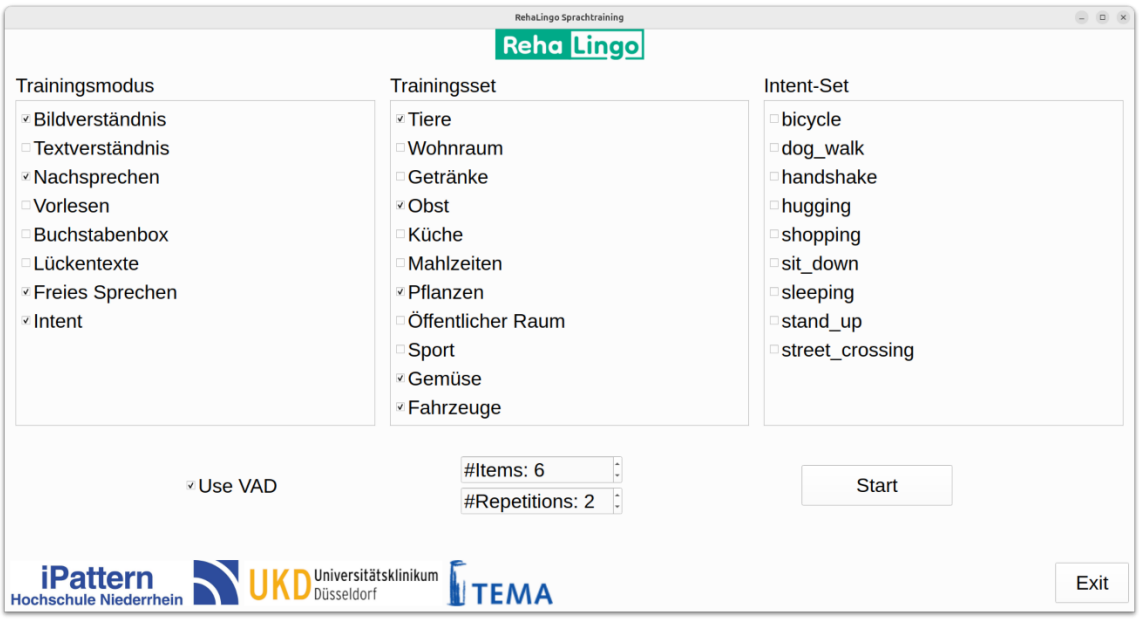

Abbildung 10: GUI zur Konfiguration eines Trainingsdurchlaufs

Zunächst werden in der linken Spalte die Trainingsmodi ausgewählt, die während eines Trainingsdurchlaufs nacheinander bearbeitet werden sollen. Die Trainingsmodi sind mit steigender Schwierigkeit gelistet, so dass sie in der Reihenfolge der Auflistung bearbeitet werden. Oben stehen die beiden Modi zum Training des Wortverständnisses. Die weiteren Modi beinhalten alle eine aktive Sprachproduktion des Patienten. Sie unterscheiden sich in der Präsentation eines Begriffs in verschiedenen medialen Formen. Beim „Vorlesen“ wird der Begriff als Bild und mit der vollständigen Textbeschreibung präsentiert. Der Term „Buchstabenbox“ steht für das in Abbildung 6 gezeigte Training. Eine ähnliche Präsentation beinhaltet der Modus „Lückentext“, mit dem Unterschied, dass keine Schriftkontur vorgegeben wird. Das „freie Sprechen“ bezieht sich auf den Modus, bei dem nur das Bild dargestellt wird, wie in den Abbildungen 7 und 8 gezeigt. Der Modus „Intent“ beinhaltet die Präsentation des Videos einer kurzen Szene, die der Patient sprachlich beschreiben soll.
In der mittleren Spalte können die Kategorien festgelegt werden, aus denen die zu übenden Begriffe für das Verständnis- und das Benenntraining ausgewählt werden. Die Auswahl mehrerer Kategorien vereinfacht das Training in der Regel. Die rechte Spalte wird nur für den Modus „Intent“ benötigt, um eine Auswahl von Szenen treffen zu können.
In dem Feld „Items“ kann die Anzahl der in einem Trainingsdurchlauf zu übenden Begriffe festgelegt werden. Es werden zufällig entsprechend viele Begriffe aus den festgelegten Kategorien ausgewählt. Damit kann die Schwierigkeit des Trainings beeinflusst werden. Im Feld „Repetitions“ kann festgelegt werden, wie häufig ein Begriff während jedes ausgewählten Trainingsmodus geübt wird.

2. Robuste Spracherkennung

Zur robusten Erkennung wurde der Ansatz betrachtet, mehrere Spracherkennungsmodule parallel einzusetzen, um die sprachliche Eingabe bei der Benutzung des Sprachtrainingssystem im Modus der Sprachproduktion zu analysieren. Abbildung 11 kann die parallele Anordnung von 3 Erkennungsmodulen entnommen werden.

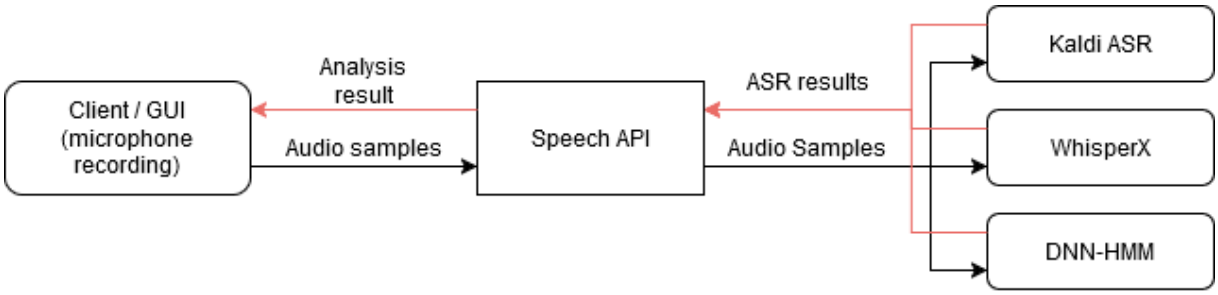

Abbildung 11: Struktureller Aufbau zur Einbindung der Spracherkennungsmodule

Mit dem parallelen Einsatz mehrerer Erkennungsmodule wird die Zielsetzung verfolgt, die individuellen Stärken der einzelnen Module zu kombinieren, um eine möglichst gute Erkennung der Spracheingaben der Aphasiker zu erreichen. Dabei ist das Ziel, die korrekte Eingabe eines Zielworts oder einen der Aphasie-Artefakte zu erkennen. Die Verwendung des am Institut für Mustererkennung entwickelten DNN-HMM [3] und des Kaldi Erkennungsmoduls [4, 5] war bereits bei der Planung des Projekts vorgesehen. Während der Laufzeit des Projekts wurde das Erkennungsmodul „Whisper“, das von der Firma OpenAI entwickelt wurde [6], verfügbar. Es wurde eingesetzt, da es eine hohe Erkennungsgüte besitzt und zudem die eingesetzten Modelle mit einer relativ kleinen Datenmenge adaptiv für eine bestimmte Anwendung trainiert werden können [7]. Im Folgenden werden die Erkennungsmodule und die notwendigen Arbeiten detailliert vorgestellt.

2.1. DNN-HMM

Dieses Erkennungsmodul besteht aus der Kombination eines tiefen neuronalen Netzes (DNN) zur Bestimmung der Emissionswahrscheinlichkeiten, die die Zuordnung eines sprachlichen Abschnitts zu einem Laut in Form eines Triphons beinhalten, und einem Hidden-Markov Modell (HMM). Es wurden eine Vielzahl von Untersuchungen zur Optimierung der Struktur und der Dimension des neuronalen Netzes mit der Zielsetzung einer möglichst hohen Erkennungsrate angestellt. Den Kern des Netzes bilden drei Bi-LSTM Schichten. Dieses System wurde zur Erkennung kleinerer Wortschätze mit bis zu einigen Hundert Wörtern konzipiert. Die Referenzmodelle zur Erkennung der Wörter werden mit Hilfe eines Aussprachelexikons durch Aneinanderhängen der entsprechenden Lautmodelle erzeugt. Die Lautmodelle wurden als Triphon Modelle unter Verwendung mehrerer Tausend Stunden deutscher Sprache trainiert. Mit einem Triphon erfolgt die Beschreibung eines Lauts in Abhängigkeit des vorhergehenden und des nachfolgenden Lauts. Mit Hilfe einer Grammatik wird die Erkennung auf bestimmte Wörter und die Reihenfolge, in der die Wörter auftreten, festgelegt und beschränkt. Damit wird die Erkennung auf den gesuchten Begriff eines gezeigten Objekts sowie Synonyme des Begriffs, phonetisch oder semantisch ähnliche Wörter oder Phrasen und alle bei Aphasikern typischerweise zu erwartenden sprachlichen Äußerungen fokussiert. Für jeden Begriff kann eine individuelle Grammatik erstellt werden. Damit kann eine gezielte Erkennung der zu erwartenden sprachlichen Eingaben erfolgen. Ein Beispiel einer solchen Grammatik ist in Abbildung 12 für den Begriff „Maus“ dargestellt. Neben der Äußerung des korrekten Zielworts oder eines Synonyms werden phonetisch und semantisch ähnliche Begriffe und Phrasen sowie Wörter einer unspezifischen Reaktion, bei der der gesuchte Begriff nicht genannt wird, in die begriffsspezifische Grammatik des Zielworts aufgenommen. Dabei können auch Wörter mit Lautfolgen aufgenommen werden, die im deutschen Lexikon nicht enthalten sind und beispielsweise durch die fehlerhafte Artikulation eines einzelnen Lauts geäußert werden. Als Beispiel ist das Wort „Mausch“ in Abbildung 12 aufgeführt. Es wurden verschiedene Ansätze entwickelt und realisiert, um solche Grammatiken automatisch durch Verwendung von semantisch aufgebauten Lexika [8, 9] und Aussprachelexika [10] zu erstellen. Zudem wurden durch das Hinzuziehen und Befragen von „Large-Language Modellen“, z.B. ChatGPT [11], semantisch ähnliche Begriffe für das jeweilige Zielwort gefunden. Insgesamt wurden für alle etwa 200 Begriffe, die in dem Trainingssystem enthalten sind, die zugehörigen Grammatiken erstellt. Dazu wurden die automatisch erstellten Grammatiken final von einer Person überprüft und eventuell modifiziert oder korrigiert, so dass

die Generierung der Grammatiken letztlich semiautomatisch erfolgte. Zur Erstellung der Grammatiken wurden erhebliche personelle Ressourcen benötigt, da auch über einen langen Zeitraum hinweg die bei Tests aufgefallenen, nicht vorhandenen semantischen Begriffe oder Phrasen ergänzt wurden.

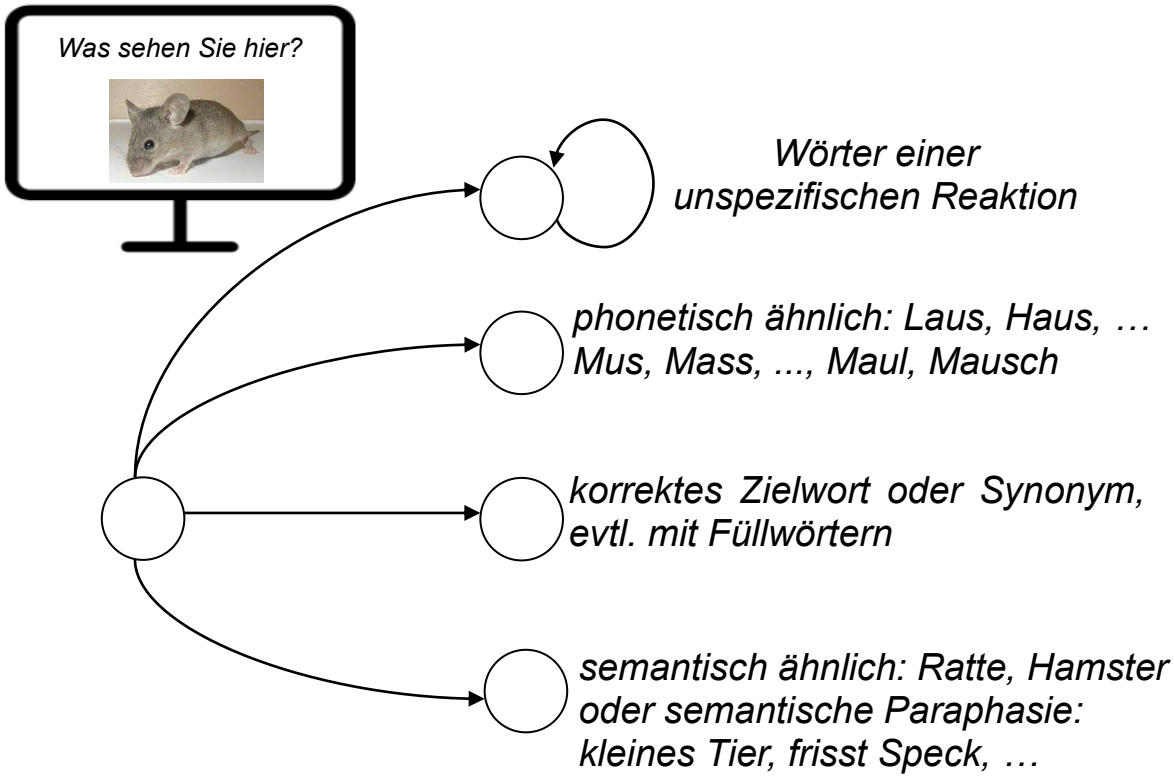

Abbildung 12: Aufbau einer zielwort-spezifischen Grammatik

2.2. Kaldi

Bei Kaldi [4] handelt es sich um ein Erkennungsmodul, mit dem im Vergleich zu dem DNN-HMM Modul große Wortschätze erkannt werden können. Kaldi wurde seit 2009 zunächst als ein frei verfügbares Werkzeug zur Forschung im Bereich der Spracherkennung entwickelt, um damit im Rahmen von Simulationen Erkennungsexperimente mit großen Sprachdatensammlungen durchzuführen. Dabei wurden neuronale Netze und das Training der Netzparameter als eine wesentliche Komponente eingebracht. Später wurde Kaldi in Kombination mit anderen Softwarewerkzeugen [12] auch zur direkten Erkennung der mit einem Mikrofon aufgenommenen Sprache eingesetzt.
Kaldi verwendet ein Lexikon, in dem in der Regel mehrere Zehntausend oder Hunderttausend Wörter lautsprachlich beschrieben sind. Im Rahmen dieses Projekts wurde ein Modell für die deutsche Sprache verwendet, das an der Uni Hamburg trainiert wurde [5]. Damit können komplette Sätze oder nahezu beliebige Wortfolgen, die die im Lexikon enthaltenen Wörter beinhalten, erkannt werden. Zur Erkennung der Sprache von Aphasikern hat dies den Vorteil, dass die Inhalte sprachlicher Äußerungen, die in keinem Zusammenhang mit dem gesuchten Begriff stehen, so genannte unspezifische Reaktionen, analysiert werden können. Dies ist mit dem DNN-HMM Modul aufgrund des eingeschränkten Vokabulars nur begrenzt möglich.

2.3. Whisper

Whisper ist ein Spracherkennungsmodul, das von der Firma OpenAI entwickelt und im Jahr 2022 veröffentlicht wurde. Es unterscheidet sich von den anderen Modulen durch die

Verwendung einer auf neuronalen Netzen basierenden Struktur, die als „Transformer“ Modell bezeichnet wird. Damit wird die zeitliche Folge von Spektren, die aus der Analyse des Sprachsignals gewonnen wird, auf eine Folge von Textzeichen abgebildet (transformiert). Es wird kein Lexikon mit einer lautsprachlichen Beschreibung von Wörtern betrachtet. Whisper erzeugt am Ausgang eine Textbeschreibung in Form einer Folge von Wörtern. Einzelne Wörter können Rechtschreibfehler beinhalten, da kein Lexikon verwendet wird und die direkte Abbildung auf eine Folge von Buchstaben erfolgt. Damit ist Whisper prinzipiell in der Lage, lautsprachliche Fehler bei der Artikulation im Text sichtbar zu machen, wobei diese Fähigkeit durch ein Training mit entsprechenden Sprachdaten, die solche Fehler beinhalten, verbessert werden könnte. Zum ursprünglichen Training von Whisper wurden etwa 680.000 Stunden Sprache verwendet. Dabei handelt es sich um die größte Datenmenge, die bis dahin zum Training eines Erkennungssystems verwendet wurde. Die Sprachdaten beinhalten verschiedene Sprachen, so dass das System direkt multilingual trainiert wurde. Zudem kann das System optional sprachliche Äußerungen in einer Sprache auf die Textbeschreibung in einer anderen Sprache abbilden.

Von OpenAI wurden verschiedene Modelle für Whisper veröffentlicht, die sich in ihrer Komplexität unterscheiden. Quantitativ unterscheiden sie sich in der Anzahl der zu trainierenden Parameter der auf neuronalen Netzen basierenden Transformer-Struktur. Mit größer werdender Zahl von Parametern steigt auch der Rechen- und Speicheraufwand bei der Anwendung des Modells, da die Anzahl der mathematischen Operationen mit größer werdender Anzahl von Parametern entsprechend wächst. Die Zahl der Parameter reicht von 39 Millionen für das „tiny“ Modell bis hin zu 1550 Millionen für das „large“ Modell. Das entspricht einem Faktor von etwa 40, was die erforderliche Rechenleistung bei Anwendung des „large“ Modells im Vergleich zum „tiny“ Modell angeht.

Eine weitere interessante Eigenschaft der Whisper Modelle ist die Möglichkeit, mit einer relativ kleinen Datensammlung ein adaptives Training für eine spezielle Anwendung durchführen zu können und damit gute Erkennungsergebnisse in der Anwendung zu erzielen [7]. Im Rahmen dieses Projekts wurde das adaptive Training des „small“ Modells mit den aufgezeichneten Sprachdaten von Aphasie Patienten betrachtet.

3. Implementierung

Im Folgenden wird vorgestellt, welche Hardware zur Realisierung des Sprachtrainingssystems gewählt wurde und wie die bereits vorgestellten GUIs und die Spracherkennungsmodule sowie die Speech-API und das „Natural Language Processing“ implementiert wurden. Zuvor wird die Einbettung des Trainingssystems als eine Komponente des gesamten Therapiesystems mit den Anteilen der von den Projektpartnern durchzuführenden Arbeiten erläutert.

3.1. Gesamtkonfiguration

In Abbildung 13 ist das gesamte Therapiesystem mit den Komponenten, die von den Projektpartnern bearbeitet wurden, dargestellt. Den Kern des Systems stellt das Sprachtrainingssystems mit den graphischen Oberflächen, die von den Patienten benutzt werden, dar. In diesem Teilprojekt wurde die robuste Spracherkennung als Basis für das Benenntraining entwickelt. Zudem wurden graphische Benutzerschnittstellen entwickelt, um mit dem Ansatz einer Dialogführung, wie sie in einer Therapiesitzung zwischen Therapeutin und Patient stattfindet, das Training durchzuführen.

Zur zentralen Speicherung und Verwaltung der Daten, z.B. den Bildern, Videos und Sprachansagen für die zu lernenden Begriffe, wird eine Datenbank verwendet. Über den gesamten Zeitraum des Projekts hinweg wurde mit den Partnern diskutiert und festgelegt, welche Daten gespeichert werden und wie der Datenaustausch mit dem Trainingssystem erfolgt. Dabei handelte es sich um einen fortwährenden Prozess, da mit der Entwicklung des Trainingssystems auch immer wieder neue Datenelemente hinzukamen. Neben Bilder-, Video- und Sprachdateien werden weitere Elemente benötigt, die beispielsweise die Konfiguration der nächsten Trainingssitzung eines Patienten oder das Protokoll und die Ergebnisse einer abgeschlossenen Therapiesitzung beinhalten. In diesem Teilprojekt wurde der Ansatz verfolgt, ein Training zu jeder Zeit und an jedem Ort zu ermöglichen, ohne dass

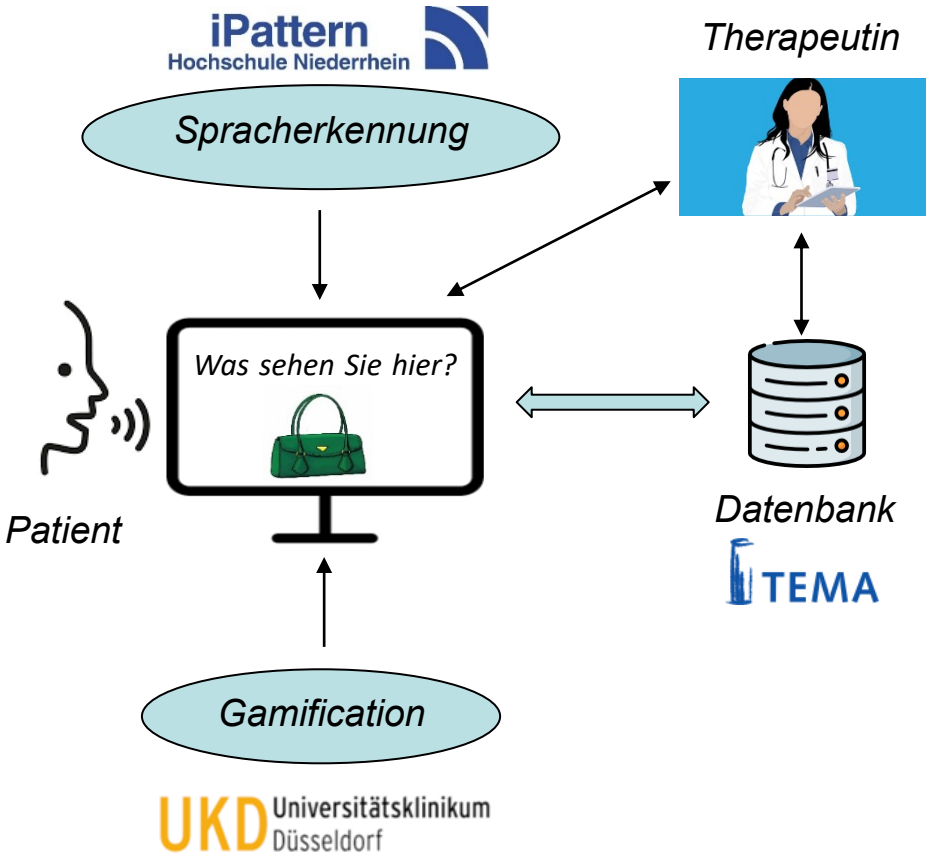

Abbildung 13: Konfiguration des gesamten Verbundprojekts

eine Verbindung zum Internet bestehen muss. Daher erfolgt der Abgleich und Datenaustausch mit der Datenbank zwischen Trainingssitzungen und in Zeiträumen, in denen eine Verbindung zum Internet besteht. Die Datei zur möglichen Konfiguration der nächsten Trainingssitzung eines Patienten, die von der betreuenden Therapeutin erstellt wurde, beinhaltet die Informationen, die in Abschnitt 1.5 bei der Vorstellung der GUI zur Konfiguration eines Trainingsdurchlaufs vorgestellt wurden. Die Log-Datei zur Protokollierung der Trainingsdurchläufe einer Therapiesitzung beinhaltet eine detailliertere Zusammenstellung, zu welchem Zeitpunkt welcher Begriff trainiert wurde und wie und mit welcher Reaktionszeit der Patient die Aufgabe bearbeitet hat. Im Fall des Benenntrainings wird die sprachliche Eingabe des Patienten und das ausführliche Erkennungsergebnis, das von der Speech-API geliefert wurde, gespeichert. Mit diesen Daten hat eine betreuende Therapeutin die Möglichkeit, bei Bedarf einen Trainingsdurchlauf oder eine ganze Sitzung zu analysieren. Die Option, einer Therapeutin die direkte Beobachtung oder die direkte Kommunikation mit einem Patienten während einer laufenden Trainingssitzung zu ermöglichen, wurde nicht betrachtet.

Am Universitätsklinikum Düsseldorf wurden die Möglichkeiten untersucht, die Motivation des Patienten durch einen spieleorientierten Ansatz („gamification“) zu steigern. Da sich dieser Ansatz deutlich von der in diesem Teilprojekt betrachteten Dialoggestaltung, die den Dialog zwischen Therapeut und Patientin nachempfindet, unterscheidet und damit auch nicht einfach integriert werden konnte, wurden am Universitätsklinikum eigene graphische Benutzer Schnittstellen für ein Sprachtraining entwickelt.

3.2. Hardware

Das Sprachtrainingssystem wurde zum Einsatz und Test mit Patienten auf 2-in-1 Laptops installiert. In Abbildung 14 ist der Aufbau zu sehen, wie er von den Patienten benutzt wurde. Mehrere 2-in-1 Laptops wurden aus den Sachmitteln des Projekts beschafft. Der 2-in-1 Laptop wurde zum Sprachtraining im „Tablet“ Modus benutzt, so dass die Bedienung über den

Touchscreen erfolgen kann.

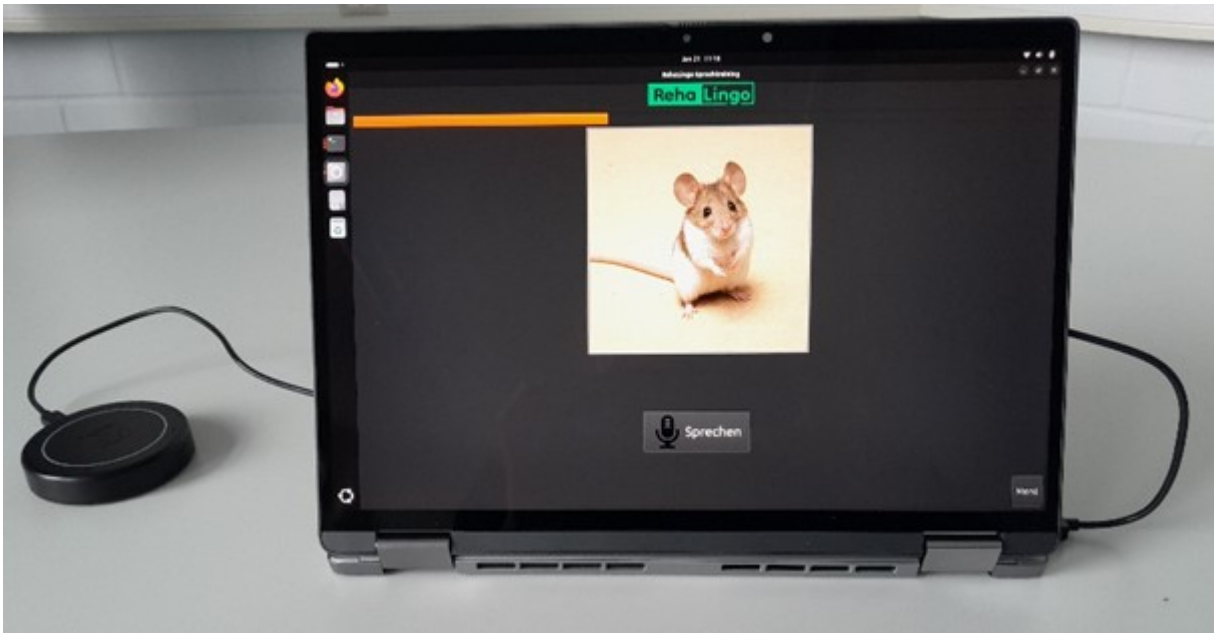

Abbildung 14: Hardwaremäßiger Aufbau des Sprachtrainingssystems

Auf dem Laptop wurde unter dem Betriebssystem Ubuntu-Linux die Software zum Training mit Hilfe der graphischen Schnittstellen installiert. Zudem sind unter Docker Compose die Container aller Spracherkennungsmodule und der selbst entwickelten Speech-API auf dem Rechner verfügbar. Der Laptop stellt eine hinreichende Rechnerleistung zur Verfügung, um die parallele Spracherkennung mit verschiedenen Erkennungssystemen ohne große Verzögerung durchführen zu können. Das entwickelte Trainingssystem kann neben dem Betrieb unter dem Betriebssystem Linux in gleicher Weise auch unter Windows verwendet werden.
Als Mikrofon wird ein per USB angeschlossenes „Respeaker“ Mikrofon der Fa. Seeed verwendet. Dabei handelt es sich um eine Anordnung von 4 Mikrofonen, deren Signale mit einem integrierten Signalprozessor verarbeitet werden. Die Signalverarbeitung des Mikrofons kann softwaremäßig konfiguriert werden. Für diese Anwendung wurde es so konfiguriert, dass eine Unterdrückung stationärer Hintergrundsignale vorgenommen wird und ein einkanaliges Ausgangssignal erzeugt wird, das für die Spracherkennung verwendet wird.
Die für das Sprachtraining erforderlichen Bilder und Videos sind lokal auf dem Rechner gespeichert, so dass ein Patient das System jederzeit und überall verwenden kann, ohne eine Verbindung zum Internet zu benötigen.

3.3. Speech-API und Erkennungsmodule

Zur Einbindung mehrerer Spracherkennungssysteme und zur Auswertung der in Textform von den Modulen gelieferten Erkennungsergebnisse wurde eine Software-Schnittstelle in Form einer Speech-API entwickelt, die in der in Abbildung 11 dargestellten Konfiguration eingebunden wurde. Die Speech-API wurde in Python programmiert. Es wurden individuelle Protokolle festgelegt, um

- die Abtastwerte des mit 16 kHz abgetasteten Sprachsignals von der GUI zur API zu übertragen,
- die Abtastwerte weiter an mehrere Erkennungsmodule zu übertragen, wobei der Transfer zu einer beliebigen Anzahl von Erkennungssystemen erfolgen kann,
- die von den Erkennungsmodulen im JSON Format bereitgestellten Ergebnisse zu empfangen und
- nach der kombinierten Auswertung aller Ergebnisse das finale Ergebnis ebenfalls im JSON Format an die GUI zu übertragen.

Die Kernkomponente der Speech-API ist ein Block zur Verarbeitung der von den Erkennungsmodulen in Textform gelieferten Ergebnisse und der Bestimmung des gegebenenfalls in der Spracheingabe enthaltenen Aphasie Artefakts. Die Auswertung der Ergebnisse mehrerer Erkennungssysteme basiert auf einem regelbasierten Ansatz, der zum einen eine Mehrheitsentscheidung beinhaltet, aber auch die speziellen Eigenschaften der Erkennungsmodule berücksichtigt. Stimmen beispielsweise die Ergebnisse verschiedener Module nicht überein, wird das Ergebnis priorisiert, dass dem Zielwort am nächsten kommt. Liefert ein Modul in diesem Fall eine Aphasie Kategorie, für deren Detektion es aufgrund seiner grundsätzlichen Eigenschaften weniger gut geeignet ist, wird dieses Ergebnis zur Auswertung nicht oder nur mit geringer Priorität berücksichtigt.
Grundsätzlich werden Wörter, die zum Ende einer Äußerung hin gesprochen werden, priorisiert gegenüber zu Beginn gesprochenen Wörtern. Auf diese Weise werden beispielsweise bei Darstellung einer „Maus“ sprachliche Eingaben wie „ähm Tier äh Ratte nein nein doch ma ma Maus“ als korrekt bewertet, obwohl sie auch zuvor semantisch ähnliche Begriffe beinhalten.
In Abbildung 15 ist auszugsweise ein JSON File dargestellt, das die Speech-API als finales Ergebnis liefert. In diesem Fall hat ein Patient für einen dargestellten „Sessel“ gemäß dem subjektiven Anhören etwas wie „sassen“ oder „sessen“ mit vorausgehendem Atemgeräusch und sprachlicher Störung im Hintergrund gesagt. Whisper hat in diesem Fall kein Erkennungsergebnis geliefert, Kaldi hat „sie saßen“ und das DNN-HMM Modul „ist was saessen“ erkannt. Kaldi ist weniger gut geeignet, um phonetisch ähnliche Äußerungen zu detektieren, die ein in einem Lexikon nicht enthaltenes Wort beinhalten. Daher besteht das finale Ergebnis in der vom DNN-HMM Modul detektierten Kategorie „phonetisch ähnlich“. Zudem werden in der Speech-API auch einige Parameter bestimmt, die z.B. aus den zeitlichen Angaben, die von den Erkennungsmodulen geliefert werden, berechnet werden. Unter dem Schlüsselwort „speed“ findet man beispielsweise die zeitliche Länge des Sprache beinhaltenden Teils der Spracheingabe („duration“) und den prozentualen Anteil dieses sprachlichen Abschnitts an der gesamten Aufnahme („temporal activity“). Zudem wird die Reaktionszeit des Patienten (“reaction time“) angegeben. Diese Parameter kann man in einer abschließenden Analyse dazu verwenden, um das Verhalten des Patienten in der gesamten Trainingssitzung zu dokumentieren und für den die Therapie begleitenden Therapeuten aufzubereiten.


```
"final": {
        …
    "word": "saessen",
    "class": "phonetic similar",
        …
},
        …
"speed": {
    "num_words": 2,
    "duration": 1.9200,
    "reaction_time": 2.85,
    "num_hesitations": 0,
    "temporal_activity": 0.3825,
    …
}
```

Abbildung 15: Auszugsweise Darstellung des JSON Files, das das finale Ergebnis der Spracherkennung beinhaltet

Zur Analyse der vom Erkennungsmodul „Whisper“ gelieferten Wortfolge für das spezielle Training der sprachlichen Beschreibung einer kurzen Handlung werden Werkzeuge des Frameworks „Rasa“ [13] eingesetzt. „Rasa“ verfolgt das Ziel, die Dialogführung eines Chatbots zu realisieren. Für die Analyse des Erkennungsergebnisses von Whisper werden nur die

Werkzeuge verwendet, um ein NLP zu trainieren und einzusetzen. Damit kann bestimmt werden, ob die erkannte Wortfolge den „Intent“, also die grundsätzlich richtige Beschreibung einer Handlung, beinhaltet. Zudem wird überprüft, ob die Wortfolge eine bestimmte „Entity“ beinhaltet. In allen im Rahmen des Trainings benutzten Szenen sind eine oder mehrere Personen enthalten. Mit dem NLP wird festgestellt, ob die Wortfolge die richtige Beschreibung der „Entity“ Person(en) als „Mann“ oder „Frau“ oder „Mann und Frau“ beinhaltet. Rasa beinhaltet ein Werkzeug, um die Bestimmung von „Intent“ und „Entity“ als Aufgabe des NLP anhand einer überschaubaren Menge von beispielhaften Wortfolgen zu trainieren.

3.4. Graphische Oberflächen

Die graphischen Oberflächen wurden zunächst unter „Matlab“ programmiert, das dazu verschiedene Werkzeuge und Bibliotheken zur Verfügung stellt. Es wurden eigene Funktionen zur zeitlich korrekten Wiedergabe der Bilder des in einer Datei enthaltenen Videos erstellt. Das war insbesondere erforderlich, um die zeitliche Synchronität bei der parallelen Wiedergabe von Videos zu erreichen, bei denen ein Video das gesamte Gesicht der Therapeutin und das andere Video nur die Lippenkontur beinhalten.
Da für den Einsatz von Matlab eine kostenpflichtige Lizenz benötigt wird und die Einbindung von Python Modulen teilweise aufwendig und problematisch ist, wurde im Verlauf des Projekts die Entscheidung getroffen, die weitere Entwicklung der GUIs unter Python durchzuführen. Dabei wird die Bibliothek PySide6 [14] verwendet, die auf der Softwareentwicklung Qt6 basiert, mit der plattformübergreifend graphische Benutzerschnittstellen entwickelt werden können. Mit PySide6 können GUIs entwickelt werden, die unter den Betriebssystemen Windows, Linux, macOS, iOS und Android lauffähig sind.

4. Sammlung von Sprachdaten

Um eine robuste Spracherkennung für die Spracheingaben von Aphasikern zu entwickeln, wird eine Sammlung von entsprechenden Sprachdaten benötigt. Damit kann die Konfiguration von Spracherkennungsmodulen im Rahmen von Simulationsexperimenten optimiert werden, um im Anwendungsfall gute Erkennungsergebnisse zu erzielen. Für Englisch gibt es eine kleine Sprachdatenbank [15]. Für Deutsch ist keine frei zugängliche Sammlung mit Sprache von Aphasikern vorhanden.
Daher wurde im Projekt eine Sammlung von Sprachdaten erstellt. Da zunächst keine Patienten zum Test des Trainingssystems zur Verfügung standen, wurden Spracheingaben von Personen ohne Aphasie aufgezeichnet. Dazu wurden zwei Herangehensweisen betrachtet. Mit der fachlichen Unterstützung von Prof. Frieg wurden typische Spracheingaben von Aphasikern in Aufnahmelisten zusammengestellt. Die zu sprechenden Eingaben wurden von 20 Personen bei Verwendung des „Respeaker“ Mikrofons, das zur Sprachaufzeichnung bei dem Trainingssystem eingesetzt wird, aufgezeichnet. Als weitere Möglichkeit, die Eingaben von Aphasikern mit den entsprechenden Artefakten nachzuempfinden, wurde eine GUI entwickelt, in der Gegenstände nur zum Teil bildlich dargestellt werden. In Abbildung 16 ist dies beispielhaft zu sehen.
Die Personen ohne Aphasie wurden instruiert, während der Aufnahmesitzung jeweils spekulativ zu äußern, was in dem Bild dargestellt ist. Dabei kamen Aufnahmen zustande, die ähnliche Artefakte wie bei aphasischer Sprache beinhalten. Insgesamt wurden mit beiden Vorgehensweisen etwa 1330 Äußerungen von Personen ohne Aphasie mit dem „Respeaker“ Mikrofon aufgezeichnet.
Im Jahr 2024 kam ein Kontakt zum LogoZentrum in Lindlar zustande, einem Therapiezentrum, das auf die intensive Sprachtherapie von Aphasikern spezialisiert ist. Das Sprachtrainingssystem konnte an 4 Terminen im Jahr 2024 und an 2 Terminen im Jahr 2025 mit jeweils 4 Patient*innen in seinem jeweiligen Entwicklungsstadium getestet werden. Mit jedem Patienten fand einzeln in einem Raum eine einstündige Sitzung statt. Sie oder er wurde dabei von einem Mitarbeiter der Hochschule betreut. Da die meisten der Patient*innen bereits in einem fortgeschrittenen Stadium ihrer Therapie waren, hatten sie nahezu keine Probleme mit den Verständnisübungen. Daher wurde der größte Teil einer Sitzung zur Durchführung von Benennübungen genutzt.

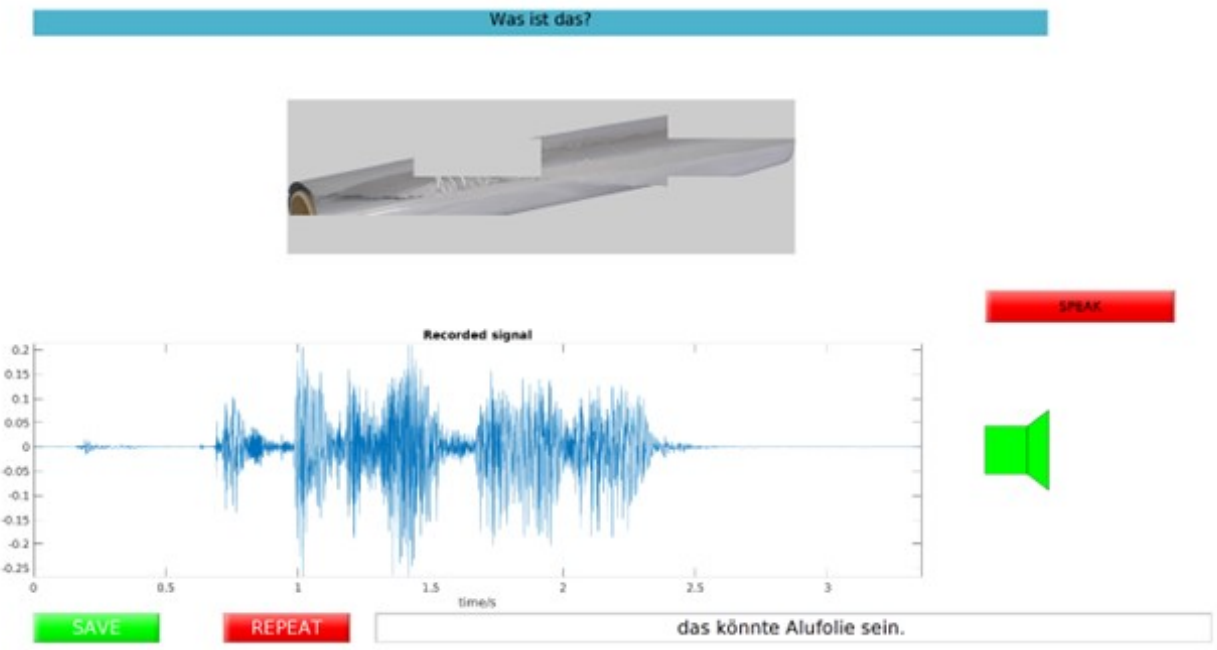

Abbildung 16: GUI zur Aufnahme von Sprachdaten

Die Patient*innen gaben ihr Einverständnis, die sprachlichen Eingaben beim Benenntraining aufzuzeichnen, um sie für spätere Erkennungsexperimente und für ein adaptives Training der Erkennungssysteme verwenden zu können. Bei den 16 Therapiesitzungen im Jahr 2024 wurden etwa 1470 Spracheingaben aufgezeichnet. Weitere 880 Äußerungen wurden im Jahr 2025 aufgenommen, so dass insgesamt 2350 Aufnahmen von Aphasikern aufgezeichnet wurden. Damit sind etwa 3680 Aufnahmen von Personen ohne und mit Aphasie vorhanden. Alle Aufnahmen wurden durch eine Person annotiert. Dabei wurde die einer Aufnahme zuzuordnende Folge von Wörtern als auch die zuordnende Klasse des Aphasie Artefakts manuell festgelegt.

5. Experimente mit den gesammelten Sprachdaten

Die aufgezeichneten Daten und die Annotationen wurden zur Durchführung von Erkennungsexperimenten genutzt. Nach der Verfügbarkeit des Whisper Erkennungssystems mit Modellen verschiedener Größe wurden Experimente zur Erkennung von 10% der gesammelten Sprachdaten gemacht. Dabei stellten sich die in Tabelle 1 aufgeführten Wortfehlerraten (WER – word error rate) ein.

Whisper Modell	Base	Small	Medium	Large	Small-adaptiert
WER / %	74.8	51.4	44.6	37.5	16.6

Tabelle 1: Wortfehlerraten mit verschiedenen Whisper Modellen

Wie zuvor erwähnt, bildet Whisper die erfasste Sprache auf eine Folge von Zeichen ab und fügt dabei Leerzeichen ein, um damit die Zeichenfolge in Wörter zu separieren. Da kein Lexikon verwendet wird, können dabei Rechtschreibfehler auftreten. Zudem werden sprachliche Artefakte möglicherweise auf Zeichenfolgen abgebildet. Daher können Wortfehler auftreten, die eigentlich keine sind und auf eine fehlerhafte Rechtschreibung oder auf sprachliche Artefakte zurückzuführen sind. Dies ist bei den in Tabelle 1 dargestellten Fehlerraten zu berücksichtigen. Die Prozentangaben werden an dieser Stelle nur herangezogen, um die verschiedenen Modelle relativ zueinander zu vergleichen. Der erwartete Effekt einer Reduktion der Fehlerrate mit einem größer werdenden Modell wird sichtbar.
Bei Experimenten mit den verschiedenen Modellen auf dem zum Training verwendeten Laptop

stellte sich heraus, dass sich bei Verwendung des „small“ Modells mit den zugehörigen Anforderungen an Rechenzeit und Speicher eine tolerable Verzögerung bis zur Verfügbarkeit des Erkennungsergebnisses einstellt. Daher wurde im Weiteren das „small“ Modell verwendet. Mit 90% der gesammelten Sprachdaten, die nicht für die Erkennungsexperimente verwendet wurden, wurde ein adaptives Training des Whisper Modells „small“ durchgeführt. Dabei stellte sich eine signifikante Reduktion der Wortfehlerrate von 51.4% auf 16.6% ein.

Für die Dialogführung des Sprachtrainingssystems ist die korrekte Erkennung des Aphasie Artefakts wichtig. Dazu wird die Einteilung in 9 Artefakt Klassen, die in der Konzeptphase festgelegt wurden, betrachtet. Die Wortfehlerrate ist daher nicht der entscheidende Parameter zur Bewertung der Erkennungssysteme, sondern die Detektionsrate des Aphasie-Artefakts. Mit etwa 150 Äußerungen, die nicht zum adaptiven Training des Whisper Modells verwendet wurden, wurde ein Experiment zur Detektion der Artefakt Klasse durchgeführt.

Erkennungsmodul	DNN-HMM	Kaldi	Wisper-small (ohne Adaption)	Speech-API	Wisper-small (mit Adaption)
Detektionsrate des Artefakts/ %	44	57	42	66	84

Tabelle 2: Detektionsraten des Aphasie Artefakts

Den in Tabelle 2 angegebenen Detektionsraten von 44% für DNN-HMM, 57% für Kaldi und 42% für Whisper kann man entnehmen, dass die Detektion des Aphasie Artefakts bei alleiniger Verwendung eines Erkennungssystems nur begrenzt funktioniert. Durch eine kombinierte Aus- und Verwertung der 3 Erkennungsergebnisse in der Speech-API kann die Detektionsrate auf 66% verbessert werden.
Durch das adaptive Training des Whisper Modells kann die Detektion deutlich auf einen Wert von 84% gesteigert werden. In Abbildung 17 ist die Verwechslungsmatrix dargestellt, die sich für die Detektion der Artefakte bei den etwa 1470 im Jahr 2024 aufgezeichneten Äußerungen ergibt.

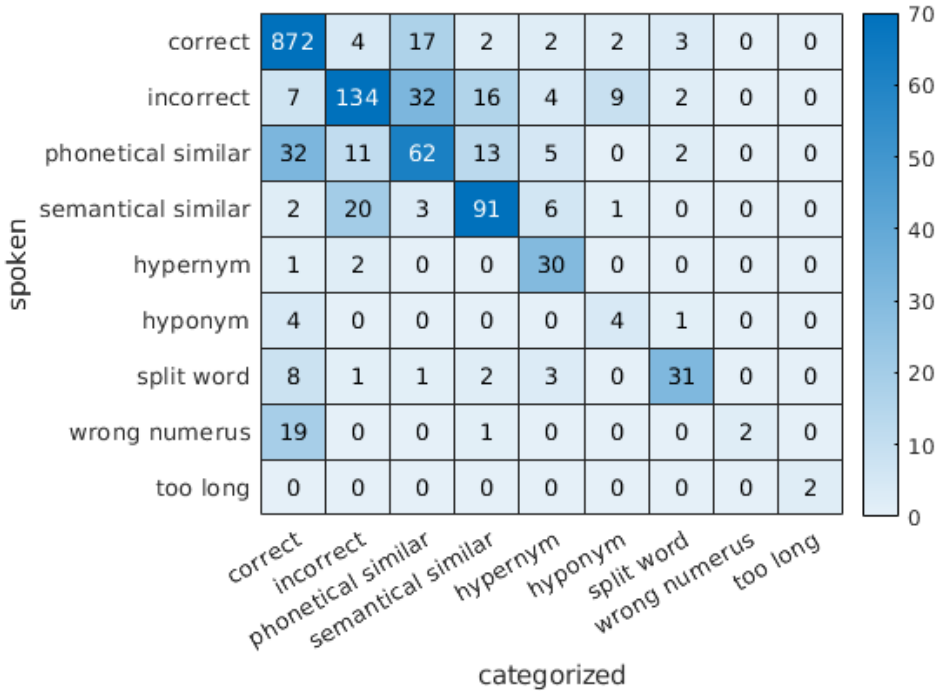

Abbildung 17: Verwechslungsmatrix bei der Detektion der Aphasie Artefakte

Große Werte, die nicht auf der Hauptdiagonalen der Matrix sind, zeigen die hauptsächlich auftretenden Fehler an. Dabei gibt es Fehler, die in Bezug auf das Sprachtraining und den Dialog mit dem Patienten weniger kritisch sind. Wird beispielsweise eine phonetisch ähnliche Eingabe als korrekt erkannt, erhält der Patient eine positive, motivierende Rückmeldung, obwohl die Eingabe nicht ganz korrekt, aber auch nicht vollständig falsch war. Wird eine semantisch ähnliche Eingabe als falsch erkannt, ist die Reaktion der Therapeutin auch nicht grundsätzlich falsch. Es fehlt nur der Hinweis, dass der Patient schon nahe am richtigen Ergebnis ist. Ebenso ist die nicht erkannte Mehrzahl (Numerus), die häufig nur durch die Nichterkennung eines angehängten „n“ oder „e“ für die Mehrzahl bedingt ist, kein kritischer Fehler. Damit spiegelt die Matrix auch die subjektive Bewertung der Mitarbeiter, die die Patienten betreuten, wider. Es treten nur wenige Fehler auf, die kritisch sind und den Patienten irritieren. Die ist z.B. der Fall, wenn der Patient einen Begriff korrekt äußert, die Therapeutin ihm aber nur eine phonetische Ähnlichkeit vermittelt („Es klingt ähnlich“).

6. Erfahrungen bei den Tests mit Patienten

Im Folgenden werden einige Erfahrungen beschrieben, die bei den Tests des Trainingssystems mit Patienten in den Jahren 2024 und 2025 gemacht wurden. Es wurde bei jedem Besuch des Therapiezentrums das System in seinem jeweiligen Entwicklungsstadium eingesetzt. Da die Patienten im Rahmen ihrer Therapiepläne in eine enge zeitliche Taktung eingebunden waren, war die Möglichkeit nicht gegeben, das System mit einem Fragebogen nach einer Sitzung schriftlich zu evaluieren. Grundsätzlich fällt es den meisten Patienten mit Aphasie ohnehin schwer, eine Evaluierung vorzunehmen. Nach einer Sitzung haben die Mitarbeiter der Hochschule jeweils versucht, die Eindrücke der Patienten in einem kurzen Gespräch zu erfragen.
Die Leitung des Therapiezentrums hat die Patienten für die Tests ausgewählt, wobei die meisten sich schon in einem fortgeschrittenen Stadium der Therapie befanden. Begonnen wurde das Training immer mit Übungen zum Verständnis gehörter Begriffe. Bis auf eine Patientin bearbeiteten alle beteiligten Probanden diese Übung zügig und nahezu fehlerfrei, so dass der überwiegende Teil einer Therapiesitzung für die Übungen zum Benennen von Begriffen verwendet wurde.
Bei den ersten Terminen wurde nur das Kaldi Modul zur Spracherkennung eingesetzt. Dabei traten Erkennungsfehler in einigen Fällen auf, wenn ein Patient das Zielwort korrekt als einzelnes Wort ohne weitere Wörter gesprochen hat. Kaldi wird zur Erkennung ganzer Sätze oder zusammenhängender Phrasen eingesetzt. Die Fehler bei der Erkennung einzelner Wörter lassen sich darauf zurückführen, dass das eingesetzte Sprachmodell nur begrenzt für diesen Fall trainiert wurde, in dem keine Zusammenhänge mit vorausgehenden Wörtern vorhanden sind. Es ließ sich beobachten, dass schon das Voranstellen eines bestimmten oder unbestimmten Artikels die Erkennung deutlich verbessert. Bei Einsatz aller 3 Erkennungsmodule traten solche Fehler nahezu nicht mehr auf. Bei den Tests an verschiedenen Terminen mit unterschiedlicher Konstellation der Erkennung waren zufällig zwei Patienten zweimal beteiligt. Sie äußerten jeweils ihren Eindruck, dass sich die Erkennung deutlich verbessert hat.
Bei den ersten Terminen wurden zudem einige Verbesserungsmöglichkeiten erkannt, was die Gestaltung und Programmierung der GUIs betraf. Es wurde beispielsweise die Möglichkeit ergänzt, dass ein Patient sich ein Video der Therapeutin mit der korrekten Benennung eines Begriffs anhören kann, wenn er selbst keine Idee hatte, wie der dargestellte Begriff zu benennen ist. Zudem trat ein unvorhergesehenes Verhalten des Systems auf, wenn ein Patient unkontrolliert mehrfach hintereinander den Bildschirm berührte, um eigentlich nur eine Aktion auszulösen.
Zur Aktivierung und Durchführung der Sprachaufnahme bei den Benennübungen wurden verschiedene Varianten getestet. Der Patient braucht nur die Fläche mit dem Mikrofonsymbol und der Beschriftung „Sprechen“ kurz zu berühren, wie es beispielsweise in Abbildung 6 dargestellt ist. Es wurden alternativ die Beendigung der Aufnahme durch eine automatische Detektion des Sprachendes mit einem VAD oder die Aufnahme über einen festgelegten Zeitraum hinweg eingesetzt. Beide Varianten erwiesen sich als einsetzbar mit den jeweiligen Vor- und Nachteilen. Bei der Aufnahme über einen definierten Zeitraum hinweg besteht die Gefahr, dass die festgelegte Zeit für die Erfassung einer längeren sprachlichen Äußerung nicht ausreicht. Zudem geht damit eventuell bei einer kurzen Äußerung eine zeitliche Verzögerung einher, bis das

Erkennungsergebnis zur Verfügung steht. Bei Einsatz eines VAD besteht die Gefahr, dass das Ende bei Störgeräuschen im Hintergrund möglicherweise nicht erkannt wird und es daher auch zu einer Verzögerung kommt. Insgesamt erwies sich die Realisierung mit einem VAD als weniger problematisch, so dass diese Variante eingesetzt wird.
Zwei Beispiel zeigten allerdings, dass dieser Start der Aufnahme durch das kurze einmalige Berühren des „Sprechen“ Bedienfelds als eine möglichst einfach gedachte Form der Bedienung für Patienten mit Aphasie problematisch sein kann. Ein Patient war es von einer anderen Anwendung zur Sprachaufzeichnung gewöhnt, das Bedienfeld bis zum Ende der Aufnahme dauerhaft zu berühren. Eine andere Patientin verstand die Aufschrift „Sprechen“ so, dass sie bei der Spracheingabe selbst „sprechen“, gefolgt von dem Begriff, äußerte. Die Spracherkennung hat dies nicht beeinflusst, da das Auftreten von Füllwörtern vorgesehen ist. Aber bei beiden Aphasie Patienten war es schwierig, ihr Verhalten auch durch mehrmalige Hinweise auf die korrekte Bedienung zu verändern.
Einige interessante Erfahrungen wurden auch bei den Übungen zur sprachlichen Beschreibung einer Handlung gemacht. Die Übung wurde von fast allen Probanden als interessant empfunden, obwohl einige aufgrund des Stadiums ihrer Therapie oder aufgrund ihrer grundsätzlichen Fähigkeiten nicht in der Lage waren, die Szene mit einer Phrase oder mehreren Wörtern sprachlich zu beschreiben.

7. Referenzen

[1] Pinkney, J.: “MTCNN Face Detection”, https://github.com/matlab-deep-learning/mtcnn-face-detection/releases/tag/v1.2.4, GitHub, 2025.
[2] Stadie, N., Hanne, S., Lorenz, A., Lauer, N., und Schrey-Dern, D.: „Lexikalische und semantische Störungen bei Aphasie“, Georg Thieme Verlag, 2019, doi:10.1055/b-006-149440.
[3] Hirsch, H.G.: “Speech Assistant System With Local Client and Server Devices to Guarantee Data Privacy”, Frontiers in Computer Science, 4, 2022, doi:10.3389/fcomp.2022.778367.
[4] Povey, D. et al.: ”The Kaldi Speech Recognition Toolkit“, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011
[5] Geislinger, R., Milde, B., und Biemann, C.: “Improved Open Source Automatic Sub-titling for Lecture Videos”, Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022), pp. 98–103. Konvens 2022 Organizers, 2022
[6] Radford, A., Kim, J.W. et al.: “Robust Speech Recognition via Large-Scale Weak Supervision”, OpenAI, arXiv: 2212.04356v1, 2022
[7] Gandhi, S.: “Fine-Tune Whisper For Multilingual ASR with Transformer”, https://huggingface.co/blog/fine-tune-whisper, 2022
[8] Hamp, B. und Feldweg, H.: "GermaNet - a Lexical-Semantic Net for German." Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, 1997.
[9] Henrich, V: und Hinrichs, E.: "GernEdiT - The GermaNet Editing Tool". Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), 2010, pp. 2228-2235
[10] Schiel, F.: „Phonolex“, Bayerisches Archiv für Sprachsignale, 2013, https://www.phonetik.uni-muenchen.de/forschung/Bas/BasPHONOLEXdeu.html
[11] OpenAI: „ChatGPT“, https://chatgpt.com/
[12] Alumäe, T.: “Full-duplex Speech-to-text System for Estonian”, Baltic HLT, 2014, doi:10.3233/978-1-61499-442-8-3
[13] Bocklisch, T., Faulkner, J., Pawlowski, N., Nichol, A.: “Rasa: Open Source Language Understanding and Dialogue Management”, 2017, arXiv: 1712.05181
[14] Python module PySide6: https://pypi.org/project/PySide6/
[15] AphasiaBank, Multimediale Datenbank mit englischen Sprachdaten und Videos von Aphasikern, https://talkbank.org/aphasia/