Richtlinie zum „Software-Sprint“

Hands-Free-Phone

Open-Source-Sprachassistent für Telefonsoftware

Schlussbericht

Zuwendungsempfänger:

Karsten Ohme

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS22S34 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Menschen mit taktilen Einschränkungen können ein Telefon nur schwer oder gar nicht bedienen. Viele Eingaben sind nötig, um einen Anruf zu tätigen anzunehmen oder einen Kontakt abzuspeichern. Die Erreichbarkeit, das Erledigen von Aufgaben und die Aufrechterhaltung von sozialen Kontakten sind deshalb stark beeinträchtigt. Das Telefon könnte aber eine der wenigen verbliebenen Möglichkeiten sein, den Kontakt zu anderen Personen aufrechtzuerhalten. Deshalb hatte das Projekt Hands-Free-Phone das Ziel einen Sprachassistenten für die Bedienung einer Telefonsoftware bereitzustellen. Dazu war es notwendig verschiedene Spracherkennungs-, Sprachauswertungs- und Sprachsynthesesysteme zu testen und zu integrieren, sowohl Online-Cloudlösungen als auch Offline-Lösungen, um einerseits Sprache in Text umzuwandeln, andererseits Sprache zu verstehen und Befehle für die Telefonbedienung abzuleiten und Sprache wiederzugeben. Dies sind Teilbereiche der künstlichen Intelligenz bzw. Machine Learning. Des Weiteren mussten verschiedene Hardwarekonstellationen aus Mini PCs und Konferenzlautsprechern evaluiert werden, um eine gute Ausführungsgeschwindigkeit und Spracheingabequalität zu erlangen.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Die 12. Runde des Prototype Funds hatte als Themenschwerpunkt unter anderen „Civic Tech“, welche nach Definition u.A. die Verbesserung des Allgemeinwohls

durch Kommunikationsplattformen ermöglichen soll. Hands-Free-Phone passt sich in diesen Themenschwerpunkt sehr gut ein.

Ausführliche Darstellung der Ergebnisse

Hands-Free-Phone stellt einen Sprachassistenten für die Bedienung einer Telefonsoftware bereit. Die Software wird auf einem Linux-Computer, vorzugsweise einem strom- und platzsparenden lüfterlosen Mini PC wie einem Raspberry Pi installiert. Auf diesem Gerät ist ein ausgewählter VoIP/SIP Anbieter konfiguriert, welcher für die Telefonsoftware verwendet wird. Am Computer angeschlossen ist ein Konferenzlautsprecher, welcher gut positioniert die Sprache aufnimmt und Umgebungsgeräusche und Echos herausfiltert.

Hands-Free-Phone besteht aus vier verschiedenen Teilkomponenten: Schlüsselworterkennung, Spracherkennung (Speech-To-Text, kurz STT), Sprachauswertung (Natural Language Processing, kurz NLP), und Sprachsynthese (Text-To-Speech, kurz TTS).

Die Schlüsselworterkennung lauscht permanent auf Audiosignale und wartet auf die Erkennung eines Schlüsselwortes. Ein Schlüsselwort ist wichtig, um die eigentliche Spracherkennung erst zu starten, wenn ein potenzieller Befehl ausgeführt werden soll (bekannte Beispiele sind „Hey, Google …“, „Alexa …“, „Siri …“).

Die Schlüsselworterkennung findet offline statt, d. h. die Erkennung findet allein auf dem verwendeten Computer statt. Als Lösung wird hier Vosk eingesetzt. Vosk ist eigentlich ein offline STT System, welches auf Kaldi basiert. Mit den verfügbaren Sprachmodellen für Deutsch hat Vosk eine schlechte Wortfehlerrate (word error rate, kurz WER), es ist aber sehr schnell und auch auf einem Raspberry Pi einsetzbar. Für das definierte Schlüsselwort „Telefon“ ist die Erkennung sehr gut.

Der nächste Teil, die Spracherkennung, ist verantwortlich dafür, die Audiosignale der Sprache in einen Text umzuwandeln. Kritisch für eine gute Spracherkennung ist auch die Qualität des aufgenommenen Audiosignals. Konferenzlautsprecher sind für eine gute Aufnahmequalität besonders geeignet, da sie mehrere Mikrofone haben, sich auf die Richtung des Sprechers einstellen können (Beamforming), Echos unterdrücken und Hintergrundgeräusche und Frequenzen herausfiltern können. Es wurden deshalb auch diverse Konferenzlautsprecher getestet und signifikante Unterschiede festgestellt. Verschiedene STT-Systeme und Konferenzlautsprecher wurden im separaten Projekt „STT Conference Speaker Test“ getestet und als Auswahl wurden die SST-Systeme von Microsoft Azure, Google Cloud, IBM Cloud, OpenAI Whisper und Coqui implementiert. Die Bewertung der Konferenzlautsprecher erfolgte in einem Kurztest. Andere Testkonstellation, wie Sprecher*in, Entfernung und Hintergrundgeräusche, können zu einem anderen Ergebnis führen. Die SST-Systeme werden ständig weiterentwickelt und es können sich auch hier Verbesserung ergeben. Von den implementierten STT-Systemen sind OpenAI Whisper und Coqui Offline-Lösungen, das heißt, dass die Spracherkennung komplett auf dem Gerät ausgeführt wird. Die WER von Coqui ist leider nicht ausreichend, aber besonders bemerkenswert

ist OpenAI Whisper, welches exzellente Ergebnisse erzielt. Es ist gleichauf oder besser als die verwendeten Cloudlösungen, OpenAI Whisper benötigt allerdings dedizierte teurere Hardware mit GPU Unterstützung, um in vertretbarer Zeit eine Transkription zu erstellen. Die Hardware NVIDIA Xavier NX wurde hierfür verwendet.

Die Sprachauswertung ist zuständig für das Erkennen der Semantik von Sätzen und die Übersetzung in auszuführende Kommandos. Der Einsatz von NLP ermöglicht nicht nur die starre Definition von Regeln zu exakten wortgenauen Phrasen, sondern auch das unscharfe Erkennen von Absichten und die trotzdem korrekte Zuordnung zu Kommandos. Das System wurde mit der NLP Lösung Rasa implementiert. Es wurde ein flexibles Vokabular für Interaktionen mit dem Telefon erstellt und die Erkennung von Telefonnummern und Namen.

Schlussendlich wurden die TTS-Systeme von Microsoft Azure, IBM Cloud, Google Cloud und Coqui TTS umgesetzt. TTS ist notwendig, um die Nutzenden über den Zustand des Systems und in Ausführung befindliche Kommandos zu informieren. Aus Text wird Sprache generiert und über den Konferenzlautsprecher ausgegeben. Das System Coqui TTS ist eine offline Lösung und liefert gute Ergebnisse, ohne zu unnatürlich zu klingen.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Die Zielgruppe des Projektes sind Menschen mit taktilen Einschränkungen, welche ein Telefon nicht mehr mit den Händen bedienen können. Darüber hinaus sollen auch allgemein Personen erreicht werden, welche durch die Sprachsteuerung einer Telefonsoftware Vorteile erhalten. Durch die Installation von Hands-Free-Phone wird dieser Zielgruppe eine bessere soziale Teilhabe ermöglicht und der Lebensalltag vereinfacht. Anhand der Veröffentlichung des Projektes als Open Source ist es möglich selbst preiswerte Installationen durchzuführen. Des Weiteren besteht für andere Entwickler die Möglichkeit Anpassungen und Beiträge an Hands-Free-Phone vorzunehmen, um das Projekt zu verbessern.

Als Weiterentwicklung wird eine Integration in bestehende Home-Automation-Systeme angestrebt. Im Open-Source-Bereich gibt es hier unter anderem die fortgeschrittene Lösungen von Rhasspy. Rhasspy erlaubt eine flexible Integration von externen SST-, TTS- und NLP-Systemen. Rasa als NLP-Lösung ist bereits integriert. Auch bietet Rhasspy eine bessere Unterstützung für Stilleerkennung am Ende eines Befehls, was in Hands-Free-Phone nur anhand eines voreingestellten Lautstärkeniveaus festgestellt wird. Defizite gibt es bei Rhasspy noch in der Schlüsselworterkennung, weil kein stimmenunabhängiges System mit hoher Erkennungsrate eingesetzt wird. Vosk wird hier angestrebt und soll integriert werden.

Eine weitere Bereicherung für den anvisierten Nutzerkreis ist die Steuerung von Fernsehgeräten. Fernsehgeräte bieten heutzutage die Möglichkeit zumindest die Tasten der Fernbedienung auch als API-Aufrufe zu steuern. Über Sprache können diese Tasten aufgerufen werden und auch Befehlsketten wie das Speichern eines

TV-Senders, das Umschalten auf einen TV-Sender und Smart-TV-Funktionen wie die Navigation eines Webbrowsers oder von Videoportalen sind angedacht.

Ebenso ist die Spracherkennungsgeschwindigkeit mit OpenAI Whipser noch merklich verzögert. Die 2023 erschienene Generation des GPU-Computers NVIDIA Orin NX soll um den Faktor 5 schneller sein, was die Offline-Erkennungsgeschwindigkeit noch schneller und praktikabler machen sollte.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Alle geplanten und im Projekt zusätzlich entstandenen Vorhaben konnten umgesetzt werden.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Das Projekt kann auf der Projektseite (https://handsfreegadgets.github.io/hands-free-phone/) und auf GitHub (https://github.com/HandsFreeGadgets/hands-free-phone) eingesehen werden.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Die veranschlagte geplante Zeit wurde moderat überschritten. Der Hauptgrund für die Überschreitung war die Integration von OpenAI Whisper und die notwendige Beschaffung und Installation der dafür dedizierten Hardware. OpenAI Whisper wurde erst innerhalb der Projektlaufzeit veröffentlicht und war in der Planung nicht absehbar. Diese Neuerung führte dazu, dass im Projekt das Thema Offline-Spracherkennung an Relevanz gewann. Eine Nichtbeachtung dieser neuen Software, hätte allerdings wichtige aktuelle Entwicklungen der Zeit missachtet und wurde deshalb priorisiert.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Marktlösungen wie der Google Assistant bieten bereits eine überzeugende umfangreiche komfortable Lösung in diesem Bereich an. Diese Lösungen nutzen aber einen Cloudservice und übermitteln die Sprachdaten dorthin. Auch wenn die Sprachdaten nicht beim Anbieter gespeichert werden, verlassen die Daten zumeist den europäischen Raum und unterstehen dann ggf. nicht mehr den europäischen Datenschutzbestimmungen. Eine Vertrauensbeziehung zu dem Anbieter muss bestehen, damit die Sprachdaten nicht für andere Dienste herangezogen werden, biometrische Profile erstellt werden oder Drittanbieter oder staatliche Organisation darauf Zugriff haben. Hands-Free-Phone bietet hier die Möglichkeit, einer kompletten Offline-Lösung bei geeigneter Hardware an, wobei keinerlei Daten an einen Anbieter übermittelt werden. Als Online-Lösung kann alternativ zumindest ein Cloudanbieter ausgewählt werden. Für

datenschutzbewusste Anwender*innen oder Institutionen, die sich an europäische Datenschutzauflagen halten müssen, kann dieses Kriterium ausschlaggebend sein, um Hands-Free-Phone einzusetzen.