Talk2Me – Sprachgesteuerte Kommissionierung mit off-the-shelf Hardware

Alfred Wertner[a], Hermann Stern[a], Franz Weghofer[b] & Viktoria Pammer-Schindler[c]
[a] Know-Center GmbH; [b] Magna Steyr; [c] Technische Universität Graz

Zusammenfassung. Sprachsteuerung stellt ein potentiell sehr mächtiges Werkzeug dar und sollte rein von der Theorie (grundlegende Spracheingabe) her schon seit 20 Jahren einsetzbar sein. Sie ist in der Vergangenheit im industriellen Umfeld jedoch primär an nicht ausgereifter Hardware oder gar der Notwendigkeit einer firmenexternen aktiven Datenverbindung gescheitert.

Bei Magna Steyr am Standort Graz wird die Kommissionierung bisher mit Hilfe von Scannern erledigt. Dieser Prozess ließe sich sehr effektiv durch eine durchgängige **Sprachsteuerung** unterstützen, wenn diese einfach, zuverlässig sowie Compliance-konform umsetzbar wäre und weiterhin den Menschen als zentralen Mittelpunkt und Akteur (Stichwort **Human in the Loop**) verstehen würde. Daher wurden bestehende Spracherkennungssysteme für mobile Plattformen sowie passende „off the shelf" Hardware (Smartphones und Headsets) ausgewählt und prototypisch als Android Applikation („Talk2Me") umgesetzt. Ziel war es, eine Aussage über die Einsetzbarkeit von sprachgesteuerten mobilen Anwendungen im industriellen Umfeld liefern zu können.

Mit dem Open Source Speech Recognition Kit CMU Sphinx in Kombination mit speziell auf das Vokabular der abgebildeten Prozesse angepassten Wörterbüchern konnten wir eine sehr gute Erkennungsrate erreichen ohne das Sprachmodell individuell auf einzelne MitarbeiterInnen trainieren zu müssen.

Talk2Me zeigt innovativ, wie erprobte, kostengünstige und verfügbare Technologie (Smartphones und Spracherkennung als Eingabe sowie Sprachsynthese als Ausgabe) Einzug in unseren Arbeitsalltag haben kann.

Das Know-Center Graz wird im Rahmen des österreichischen COMET-Programms – „Competence Centers for Excellent Technologies" unter der Schirmherrschaft des Bundesministeriums für Verkehr, Innovation und Technologie (BMVIT), des Bundesministeriums für Digitalisierung und Wirtschaftsstandort (BMDW) und des Landes Steiermark gefördert. Das COMET-Programm wird von der Österreichischen Forschungsförderungsgesellschaft (FFG) verwaltet.

1. Einleitung

Technologie umgibt uns! Wir haben ständig ein Smartphone zur Hand. Wir zeichnen unsere Fitness und unseren Schlaf auf. Wir spielen mit virtuellen Brillen, gehen ins 3D Kino und fragen Siri oder Google, wo es den nächsten Coffee2Go gibt. All diese Technologien haben in einer Geschwindigkeit und in einer Dimension Einzug in unser (Privat)Leben gehalten, wie wir es nie für möglich gehalten hätten.

Spracherkennung beispielsweise stellt ein potentiell sehr mächtiges Werkzeug dar und sollte rein von der Theorie (grundlegende Spracheingabe) her schon seit 20 Jahren einsetzbar sein. Dennoch tut sich die Industrie sehr schwer, diese Technologien für ihre Zwecke sinnvoll zu adaptieren. Idealerweise wird **Spracherkennung** gleich mit **Sprachsynthese** (auch die Antwort erfolgt mittels gesprochener Sprache) gekoppelt, um eine durchgängige Bedienung (**Sprachsteuerung**) über Sprache zu ermöglichen. Um gesprochene Sprache zu „verstehen", muss ein System gesprochene Worte aufzeichnen und mittels Signalverarbeitung, Machine Learning und einem darunterliegenden Sprachmodell in einen computerlesbaren Text umwandeln; je begrenzter das zu erkennende Vokabular sein kann, desto besser.

2. Problemstellung

Bei Magna Steyr am Standort Graz wird die Kommissionierung (das ist das Zusammenstellen von bestimmten Teilmengen aus einer bereitgestellten Gesamtmenge aufgrund von Aufträgen) von Autoteilen bisher mit Hilfe von Scannern erledigt. Dieser Prozess ließe sich sehr effektiv durch eine durchgängige **Sprachsteuerung** unterstützen, wenn diese einfach, zuverlässig und kostengünstig umsetzbar wäre und weiterhin den Menschen als zentralen Mittelpunkt und Akteur (Stichwort **Human in the Loop**) verstehen würde.

Sprachsteuerung ist in der Vergangenheit im industriellen Umfeld jedoch primär an nicht ausgereifter Hardware oder der Notwendigkeit einer aktiven Datenverbindung (eventuell sogar zu einem firmenexternen Cloud Service wie Google) gescheitert. Dennoch sollen Unternehmensprozesse sowohl im Unternehmen selbst als auch beim Kunden vor Ort sprachgesteuert unterstützt werden können. Daher war ein Schlüsselkriterium für die Umsetzung, die **Sprachverarbeitung direkt auf dem Gerät durchführen** zu können. Von AnwenderInnen gesprochene Sprache soll offline, das heißt direkt auf dem Smartphone verarbeitet und in handlungsrelevante Texte übersetzt werden. Bedingt durch das **prozessbezogene Spezialvokabular** muss ein **Training** des dahinterliegenden Sprachmodells möglich sein. Des Weiteren soll diese Sprachsteuerung mit **handelsüblichen Smartphones und damit gekoppelten Headsets** („off the shelf" Komponenten) im industriellem Umfeld umgesetzt werden können. Die Vorteile der Verwendung von handelsüblicher Hardware liegen auf der Hand: Unternehmen sind damit unabhängig von Hardwareherstellern, die Geräte selbst sind erprobt, mit ausreichend Akkulaufzeit ausgestattet und günstig erhältlich.

Die **Analyse / Auswahl von bestehenden Spracherkennungssystemen** für mobile Plattformen sowie die Auswahl der **passenden Hardware** (Smartphones und Headsets) waren daher zentrale Erfolgsfaktoren. Darüber hinaus sollten obig als geeignet identifizierte Komponenten als **Prototyp** („Talk2Me“) umgesetzt werden, um eine Aussage über die Einsetzbarkeit von sprachgesteuerten mobilen Anwendungen im industriellen Umfeld liefern zu können.

3. Spracherkennungssysteme für mobile Plattformen

MitarbeiterInnen sollen möglichst effektiv unter Beibehaltung der notwendigen Compliance (Sprache offline verarbeiten, Training des Sprachmodells, etc.) direkt im Arbeitsprozess unterstützt werden. Dazu wurde eine Reihe von am Markt verfügbaren **Spracherkennungssystemen** in Hinblick auf die in Tabelle 1 aufgelisteten Systemanforderungen hin untersucht.

Offline	Das System unterstützt die Verarbeitung von Sprache (Erkennung sowie Synthese) am Gerät ohne Internetzugang.
Training	Das System unterstützt das Trainieren der Spracherkennung (Einsprechen von Trainingsdaten und Erzeugung eines neuen, adaptierten Akustikmodells für die Spracherkennung).
Verfügbarkeit	Das dem System beilegende Software Development Kit (SDK) und dessen Eigenschaften: die der SDK zugewiesenen Lizenz und die Kompatibilität der SDK mit gängigen mobilen Betriebssystemen.
Installation	Aufwand für die Installation der für den Betrieb notwendigen Systemkomponenten.
Integration	Aufwand für die Inbetriebnahme des Systems.

Tabelle 1: Systemanforderungen an Spracherkennungssysteme für mobile Plattformen

Google Voice Search[1]

Bei Google Voice Search ist die Offline Verarbeitung zwar grundsätzlich möglich, durch die schlechte Erkennungsrate und durch die Tatsache, dass es keine Möglichkeit gibt, das Spracherkennungssystem in irgendeiner Form zu erweitern oder verbessern, nicht zu empfehlen.

[1] https://developer.android.com/reference/android/speech/package-summary.html

Offline	Ein Offline Modus für Google Voice Search ist grundsätzlich vorhanden. Im Vergleich zu Googles Online Erkennung, die sehr gut funktioniert und auch bei einem Netzzugang mit mäßiger Geschwindigkeit gefühlt sehr schnell antwortet, verschlechtert sich das Ergebnis der Offline Variante drastisch. Die Spracherkennungsrate sinkt unter ein noch benutzbares Level. Selbst einfache Wörter wie „Birne“ oder die Zahl „6“ werden nicht richtig erkannt.
Training	Google setzt vor allem auf die Nutzung des hauseigenen Online Spracherkennungssystems; die Offline Version bietet keine Möglichkeit für ein Training. Auch andere Teile des Spracherkennungssystems wie das Wörterbuch oder die Grammatik können nicht geändert werden.
Verfügbarkeit	Die SDK ist Open Source und steht unter Apache 2.0 Lizenz zur Verfügung. Die SDK steht für Android zur Verfügung, die Kompatibilität mit anderen Plattformen ist als gut zu bewerten, da es für verschiedenste Programmiersprachen (z. B.: HTML5, Chrome Web Speech) Plug-Ins gibt. Dennoch muss für jede Plattform das geeignete SDK gesucht und individuell programmiert werden.
Installation	Um die Offline Version von Google Voice Search zu nutzen, muss lediglich das gewünschte Offline Sprachpaket heruntergeladen werden (ca. 15 - 25Megabyte je nach Sprache).
Integration	Es sind nur wenige Zeilen Code notwendig um die Funktionen von Google Voice Search zu nutzen.

Tabelle 2: Funktionsüberblick Google Voice Search

CMU Pocketsphinx[2]

Die Steuerung der Spracherkennung funktioniert sehr gut und ist individuell konfigurierbar, wodurch auch eine problemlose „always-on“ Spracherkennung möglich ist. Die Erkennung reagiert dank vollständiger Verarbeitung am Gerät sehr schnell. Aufgrund der vielen Anpassungsmöglichkeiten ist Pocketsphinx sicherlich eine gute Wahl für Ablaufsteuerungen mit einzelnen, vorgegebenen Befehlssequenzen. Sofern es nicht möglich ist Befehle zu verwenden, die sich möglichst gut voneinander unterscheiden, ist ein Training des Sprachmodells möglich und empfehlenswert. Durch den Open-Source Charakter ist ein größerer Einarbeitungsaufwand erforderlich.

[2] https://cmusphinx.github.io/wiki/tutorialpocketsphinx/

Offline	Bietet eine Spracherkennungspipeline ohne notwendigen Netzzugang.
Training	Training ist durch diverse Tools möglich. Wörterbuch oder die Grammatik können einfach geändert werden.
Verfügbarkeit	Open-Source Projekt und steht unter der BSD Lizenz zur Verfügung. Pocketsphinx ist in C geschrieben und kann dadurch auf vielen Systemen verwendet werden (Linux, Windows, MacOS, Android). Es steht ein SDK für die Android Plattform zur Verfügung.
Installation	Die Installation auf Android Geräten ist umfangreicher und benötigt etwas Einarbeitungszeit. Neben der SDK müssen auch noch geeignete Grammatikdateien und Wörterbücher für jede Sprache gesucht und heruntergeladen werden (ca. 15-100 Megabyte je nach Sprache).
Integration	Der grundsätzliche Aufbau und Verwendung ist am besten mittels der verfügbaren Demoapplikation für Android ersichtlich.

Tabelle 3: Funktionsüberblick CMU Pocketsphinx

Microsoft Speech Plattform[3]

Die Microsoft Speech Plattform bietet keine Unterstützung für die Offlineverarbeitung und die Spracherkennung ist nicht ausreichend adaptierbar (Sprachmodell ist nicht trainierbar).

Offline	Die Spracherkennung für Windows Phone baut auf Cortana auf und ist im Moment nur Online über Microsoft Server möglich.
Training	Das Akustikmodell selbst ist vorgegeben und kann nicht verändert werden. Bei Tests in normaler bis ruhiger Umgebung konnte eine mittelmäßige Erkennung vordefinierter Wörter erreicht werden. Deshalb sollte zusätzlich eine eigene Grammatik verwendet werden, da sonst die Erkennung noch etwas unterhalb der schlechten Erkennungsrate der Offline Version von Google Voice Search liegt.
Verfügbarkeit	Die SDK steht nur für Windows Plattformen zur Verfügung.
Installation	Um die Offline Spracherkennung zu nutzen sind zumindest die Sprachpakete zu installieren.

[3] https://msdn.microsoft.com/en-us/library/office/hh361572(v=office.14).aspx

Integration	Die Erstellung einer C-Sharp Anwendung ist sehr unkompliziert in wenigen Schritten möglich. Die SDK ist sehr gut dokumentiert und erlaubt einige Anpassungen. Des Weiteren ist sie durch die Verwendung der Windows Sprachpakete relativ klein.

Tabelle 4: Funktionsüberblick Microsoft Speech Platform

Nuance Dragon Naturally Speaking[4]

Das Dragon Mobile SDK Spracherkennungssystem für mobile Geräte bietet keine Unterstützung für die Offlineverarbeitung. Sämtliche Übersetzungen werden über die Nuance Cloud-Services abgewickelt und es entstehen "pro Übersetzung" Kosten. Es gibt keine Möglichkeit einen lokalen Nuance Server zu betreiben. Die Adaption der Spracherkennung ist möglich.

Offline	Eine Offline Verarbeitung ist nicht vorgesehen. Smartphones dienen als Sprachquellen, die Audiodaten an Server übermitteln.
Training	Die SDK bietet eine vorgefertigte Trainingsfunktion auf Benutzerbasis. Es wird ein vordefinierter Text aufgenommen und damit das Akustikmodell adaptiert.
Verfügbarkeit	Die Software ist nicht frei verfügbar. Es sind Lizenzgebühren für die SDKs zu bezahlen (Kosten zwischen 99€ und >1000€ pro Lizenz).
Installation	Unterstützt werden nur Windows Plattformen, seit kurzem gibt es jedoch auch ein SDK für alle gängigen mobilen Plattformen.
Integration	Ein Test der Dragon Mobile SDK war zu diesem Zeitpunkt nicht möglich.

Tabelle 5: Funktionsüberblick Nuance Dragon Naturally Speaking

4. Anforderungen an “off –the-shelf” Hardware

Im Zuge der Recherche und Tests der Spracherkennungssysteme wurden sowohl **Bluetooth** als auch **kabelgebundene Headsets** verwendet. Die Verwendung von Bluetooth Headsets zeigte dabei eine deutliche Verschlechterung der Leistung der Spracherkennung, da durch zeitliche Verzögerungen der gesprochene Text abgeschnitten wurde. Weiters konnten wir feststellen, dass ein Headset ohne Mikrofonarm die Leistung der Spracherkennung drastisch verschlechtert, weshalb wir ein kabelgebundenes Headset mit Mikrofonarm bevorzugen. Ein **geeignetes Smartphone** sollte folglich daher einen Stecker für ein Headset und aktuell übliche Leistungsdaten (Quad Core Prozessor, etc.) aufweisen.

[4] https://www.nuance.com/dragon.html

5. Talk2Me: Sprachbasierte Kommissionierung von Autoteilen mit off- -the-shelf Hardware

Die Android App **Talk2Me** wurde vom Know-Center gemeinsam mit Magna Steyr entwickelt; Ziel war es, mit handelsüblichen, erprobten und durch MitarbeiterInnen akzeptierten Technologien (einem Smartphone und einem damit gekoppelten Headset) den Einsatz von sprachgesteuerten Anwendungen im industriellen Umfeld erfolgreich zu demonstrieren und weiter voran zu treiben.

Als erstes Einsatzgebiet von Talk2Me wurde die Kommissionierung von Autoteilen ausgewählt; bisher wird dieser Prozess mit Hilfe von Scannern durchgeführt. Talk2Me hilft der MitarbeiterIn durch eine komplett sprachgesteuerte Führung (sowohl Spracherkennung als Eingabe als auch Sprachsynthese als Ausgabe) beim Zusammenstellen von Bestellungen aufgrund von Aufträgen. Damit hat die MitarbeiterIn jederzeit die Hände frei, was beispielsweise bei großen Bestellteilen oder dem Tragen von Handschuhen von Vorteil ist.

Talk2Me setzt sich aus Smartphone, Headset und der mobilen Applikation für die sprachbasierte Kommissionierung zusammen. Als Smartphone wurde aufgrund des sehr guten Preis Leistungsverhältnisses (rund 250 Euro) das One Plus One[5] ausgewählt. Das Smartphone verfügt über einen Quad Core 2.5 GHz Krait 400 Prozessor, 3 Gigabyte RAM, einem 5.5 Zoll Touchscreen und einer Audiobuchse für kabelgebundene Headsets. Als Headset wurde das Sennheiser PC 21-II[6] (rund 25 Euro) verwendet. Es ist ein kabelgebundenes Headset und hat einen vergleichsweise langen Mikrofonarm. Der lange Arm hat den Vorteil, dass das Gesprochene klarer von Umgebungsgeräuschen getrennt werden kann.

Aus der Analyse existierender Spracherkennungssysteme ging hervor, dass das Open Source Speech Recognition Kit **Pocketsphinx** die Anforderungen an eine sprachgesteuerte mobile Applikation am besten erfüllt (siehe Kapitel 3). Pocketsphinx ist das einzige System, das die beiden Hauptanforderungen, nämlich die **Verarbeitung der Sprache (Erkennung, Synthese) am Gerät ohne Netzzugang** und das **Trainieren der Spracherkennung** (Einsprechen von Trainingsdaten und Erzeugen eines neuen, adaptierten Akustikmodells für die Spracherkennung), erfüllen kann. Darüber hinaus ist Pocketsphinx mit Abstand das flexibelste System und Open Source. Die Lizenzbestimmungen erlauben eine Verwendung des Originals oder einer Ableitung davon ohne Bedingungen außer dem Vermerk des Autors in der Ableitung.

5.1 Vokabular und Sprachmodell

Für die Spracherkennung sind das zu erkennende Spezialvokabular und die dem Vokabular zugrunde liegende Sprache zu definieren. Das Vokabular für einen Kommissionierungsprozess (Auszug siehe Tabelle 7) setzt sich aus den Autoteilen, den Fächern („Number 1“ bis „Number 9“) und speziellen Wörtern zur Ablaufsteuerung (zum Beispiel "Cancel" oder

[5] https://oneplus.net/at/one
[6] https://en-us.sennheiser.com/voip-headset-skype-noise-cancelling-microphone-pc-21-ii

"Next") zusammen. Reale Bezeichnungen von Autoteilen sollten auf ein spezielles (insbesondere kürzeres und leichter memorierbares) Vokabular abgebildet werden. Aus einem "Sportlenkrad Schwarz" kann zum Beispiel "Apple" oder aus "Schaltwippe" "Red" werden (siehe Tabelle 6 mit Beispielen zu diesem Mapping).

Bezeichnung	Mapping
Standardlenkrad Schwarz	Grape
Sportlenkrad Schwarz	Apple
Komfortlenkrad Schwarz	Marille
Schaltwippe	Red

Tabelle 6: Mapping von realen Autoteilebezeichnungen auf das Vokabular in der Kommissionierung

Continue	Part Lemon	Number One
Stop	Part Red	Number Two
Cancel	Part Yellow	Number Three
Next Pick	Part Blue	Number Four
Yes	Part Apple	Number Five
	Part Banana	Number Six
Part Strawberry	Part Raspberry	Number Seven
Part Orange	Part Green	Number Eight
Part Grape		Number Nine

Tabelle 7: Vokabular für den Kommissionierungsprozess

Um das gesprochene Vokabular zu erkennen und in Text umzuwandeln, benutzt Pocketsphinx ein Sprachmodell. Dabei kann grundsätzlich auf bereits existierende Sprachmodelle zurückgegriffen werden. Für Pocketsphinx stehen Sprachmodelle für zwölf Sprachen zur Verfügung, darunter beispielsweise Amerikanisch Englisch, Deutsch, Spanisch, Russisch und Französisch.

Das Sprachmodell für Amerikanisch Englisch beinhaltet die umfangreichste Menge an Trainingsdaten und zeichnet sich durch eine gute, robuste Erkennungsrate aus. Aus diesem Grund wurde Amerikanisch Englisch als Sprachmodell ausgewählt.

5.2 Sprachgesteuerter Kommissionierungsprozess

Der klassische Vorgang der Kommissionierung basiert auf der Auftragsannahme an einem festen Ort und der manuellen Zusammenstellung von Produkten im Lager unter Verwendung proprietärer Hand-Scanner. Durch Talk2Me entfallen die Wege der Auftragsannahme, da diese mobil erfolgen kann, und die MitarbeiterIn hat während der Arbeit beide Hände frei, da die Interaktion rein über Sprache und Hören erfolgt.

Dies ist insbesondere bei der Kommissionierung sperriger Teile (beispielsweise Autotüren, siehe Abbildung 1) zu empfehlen.

Abbildung 1: Talk2Me führt durch den einen Kommissionierungsprozess

Um eine Kommissionierung erfolgreich durchzuführen muss die MitarbeiterIn Autoteile in die Fächer zweier Regale einordnen. Die Kommissionierung eines Teils läuft dabei immer nach dem gleichen Schema ab: Talk2Me spricht vor, welches Autoteil zu nehmen ist und wie viele Stück davon. Die MitarbeiterIn nimmt die richtigen Teile in der geforderten Stückzahl und bestätigt mittels Sprachkommando. Kann Talk2Me die Spracheingabe dem Vokabular zuordnen, gibt Talk2Me das Fach zum Einordnen vor. Auch hier wartet Talk2Me auf die gesprochene Bestätigung der MitarbeiterIn und sofern die Eingabe verstanden wurde, ist die Kommissionierung für diesen Teil abgeschlossen, und der nächste Teil ist an der Reihe.

Falls Talk2Me eine Eingabe nicht versteht, kann jederzeit auf Eingabe über QR Code (mittels der eingebauten Kamera) umgeschaltet. Der Prozess ist somit nicht unterbrochen. Abbildung 2 zeigt exemplarisch die Kommissionierung des Autoteils "Green" (Schaltwippe Leder) in das Fach sieben.

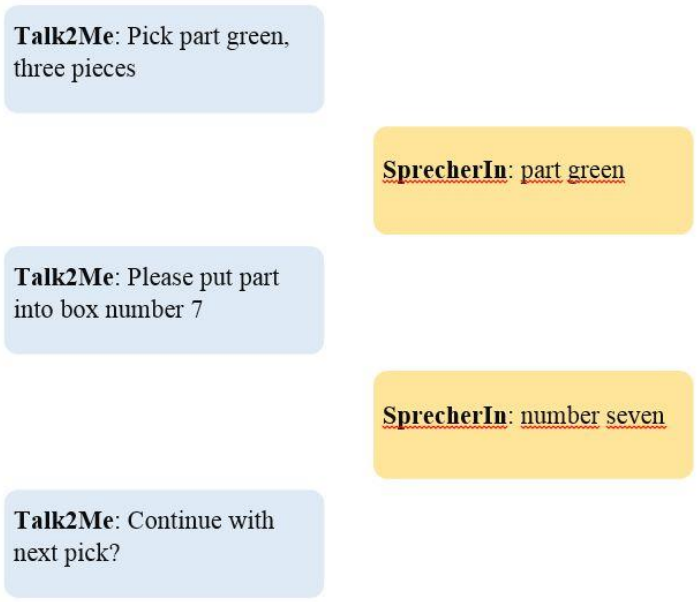

Abbildung 2: Auszug sprachgesteuerte Kommissionierung für den Autoteil "Green" (Schaltwippe Leder)

5.3 Training des Sprachmodells

Die Qualität der Spracherkennung hängt in erster Linie vom Akustikmodell des bestehenden Sprachkorpus ab. Abhängig davon und der Stimme der SprecherIn werden bestimmte Spracheingaben besser oder schlechter erkannt. Um die Trefferquote der Eingaben bei unterschiedlichen Stimmen, Tonfällen, Dialekten, etc. zu verbessern, bietet Pocketsphinx die Möglichkeit, das Sprachmodell zu trainieren und sich so an die akustischen Eigenschaften der Stimme einer Person anzupassen.

Es besteht auch die Möglichkeit, sogenannte „Noise Dictionaries“ anzulegen, mit denen Störgeräusche gezielt als Noise deklariert werden können.

5.4 Talk2Me Demonstrationskoffer

Ein Talk2Me Demonstrationskoffer veranschaulicht den Kommissionierungsprozess der Autoteile und erlaubt eine Präsentation der Talk2Me App ohne großen Aufwand (siehe Abbildung 3).

Autoteile werden dabei durch Holzklötze repräsentiert und sind mit den im Vokabular enthaltenen Farben und Obstsorten markiert. Die zwei Regale sind mit blauen und roten Feldern im Koffer hinterlegt in denen sich die einzelnen Fächer befinden. Abgesehen davon sind die verwendete Hardware sowie die mobile Applikation ident.

Abbildung 3: Talk2Me Demonstrationskoffer

6. Fazit und Ausblick

Talk2Me zeigt innovativ, wie erprobte, kostengünstige und verfügbare Technologie (Smartphones und Spracherkennung als Eingabe sowie Sprachsynthese als Ausgabe) Einzug in unseren Arbeitsalltag haben kann.

Dabei kann die komplette Verarbeitung von Sprache (Steuerung, Sprachsteuerung und Spracherkennung) auf einem handelsüblichen Smartphone und damit gekoppelten Headset durchgeführt, sowie durch Trainieren des Sprachmodells eine hohe Trefferquote erzielt werden. Zu den vielen Vorzügen von Talk2Me zählen unter anderem ein leicht an neue Prozesse adaptierbares Vokabular sowie die Möglichkeit, das verwendete Sprachmodell durch ein einfaches Training (Vorlesen von 50 Trainingssätzen) sehr individuell und damit sogar an Dialekte anpassen zu können.

Bei Talk2Me steht der Mensch weiterhin im Mittelpunkt. Er kann jederzeit den Kommissionierungsprozess selbst steuern, ist durch keine zusätzlichen Geräte (wie VR Brillen) behindert, und kann auch jederzeit auf den bisher bestehenden Ablauf per QR Code Scannen zurück wechseln. Dies wird auch dann angeboten, wenn ein Sprachbefehl doch einmal nicht verstanden werden kann. Der Prozess wird somit nicht unterbrochen.

Für die Verwendung von Talk2Me kann auf gängige, robuste und kostengünstige Android kompatible Hardware zurückgegriffen werden (und das Headset sogar selbst durch die MitarbeiterIn nach Tragekomfort gewählt werden). Zudem werden kabelgebundene Headsets verwendet, welche eine höhere Audioqualität, keine Akkuprobleme sowie geringe Strahlenbelastung als Bluetooth oder WiFi aufweisen.

Mit dem Open Source Speech Recognition Kit CMU Sphinx in Kombination mit speziell an das Vokabular der abgebildeten Prozesse angepassten Wörterbüchern konnten wir eine sehr gute Erkennungsrate erreichen – und das auch bei bestehenden Hintergrundgeräuschen und ohne das Sprachmodell individuell auf die MitarbeiterIn trainieren zu müssen. Dennoch sinkt die Trefferrate der Spracherkennung bei sehr lauten Hintergrundgeräuschen unter ein benutzbares Niveau, da das Sprachmodell unter Laborbedingungen (eher ruhige Umgebung, lautes und deutliches Einsprechen des Vokabulars) erstellt wurde.

Um das Spracherkennungssystem auch in sehr lauten Arbeitsumgebungen einsetzen zu können, wäre als nächster Schritt die Robustheit bei lauten Hintergrundgeräuschen zu verbessern. Ein Ansatz dazu ist es, Trainingsdaten auch direkt dort aufzunehmen, idealerweise werden Aufnahmeorte gewählt, an denen die Sprachsteuerung angewendet wird. So besteht eine gute Chance, Störgeräusche zu identifizieren (zum Beispiel Hämmern oder Bohren) und die Spracherkennung weiter zu verbessern.