Hallo Kaffeemaschine!

© AdobeStock_230582026_AdobeStock_230582026

Wer schon einmal im ICE im ersten Abteil hinter der Führerkabine gesessen hat, hat vielleicht die Meldung „Zugbeeinflussung! Zugbeeinflussung!“ gehört. Damit wird die Wahrnehmung des Lokführers, der bereits haptisch und visuell mit dem Zug kommuniziert, um eine weitere, akustische Ebene erweitert, die seine unmittelbare Aufmerksamkeit erreicht.

Interaktion per Sprache

Zuerst als nette Spielerei betrachtet, wurde diese im Smart Home fester Bestandteil: Die Steuerung von Musik, Licht, Erinnerungs-Timern und das Erstellen von Einkaufslisten ist mit dem Medium Sprache einfach und bequem. Während die Sprachbedienung anfänglich „nur“ einen ähnlichen Komfortgewinn wie die drahtlose Fernbedienung des TV-Geräts bot, ist mittlerweile eine Infrastruktur entstanden, in der sie einen echten Mehrwert bietet. Amazon mit Alexa als Vorreiter unterstützt die Entwicklung von Spracherkennung. In dem neuen, MASSIVE genannten Projekt stellt Amazon Datensätze in 51 Sprachen zur Verfügung, auf die Entwickler zurückgreifen können, um ihre Algorithmen und Systeme einem Test zu unterziehen.

Bedeutung der Sprachtechnologie

Dem Medium Sprache mit dem gesprochenen Wort als Eingabekommando und der synthetisierten Sprache als Ausgabe spricht man einen festen Platz neben dem traditionellen Display- und Touchscreen-Interface zu. Das Consulting-Unternehmen Gartner erstellt Studien für die Zukunft verschiedener Technologien. Der so genannte „Gartner Hype Cycle“ stellt dabei die Lebensphasen einer Technologie in fünf Stufen dar, die von der anfänglichen Euphorie über die Ernüchterung bei der Realisierung bis hin zum produktiven Einsatz reichen. Die Spracherkennung hat bereits die Phase der Produktivität erreicht; auf einem guten Wege dahin ist die Sprachsynthese. Noch Entwicklungsarbeit ist in das Verstehen und der Interpretation natürlicher Sprache zu legen.

KI-unterstützte Spracherkennung

Eine hohe Bedeutung nimmt nicht die rein algorithmische, sondern die durch Künstliche Intelligenz (KI) unterstützte Spracherkennung ein. Doch was brauchen wir für den Einsatz in einem professionellen Umfeld, anders als die gängigen Sprachassistenten, die man auch einmal bitten kann, einen Witz zu erzählen? Im Sinne eines „guten“ HMI, das ergonomisch designt ist, erwartet man eine sprecherunabhängige Erkennung des gesprochenen Wortes, möglichst in mehreren Sprachen, genau hinzuhören und auch wegzuhören (manchmal wird die Sprachbedienung getriggert, wenn das Schlüsselwort fälschlich erkannt wird), und tolerant bezüglich der Grammatik zu sein. Füllwörter wie „bitte“, „einmal“, „ja, genau“ und Räuspern sollen bitte ignoriert werden und nicht zu Fehlbedienungen führen.

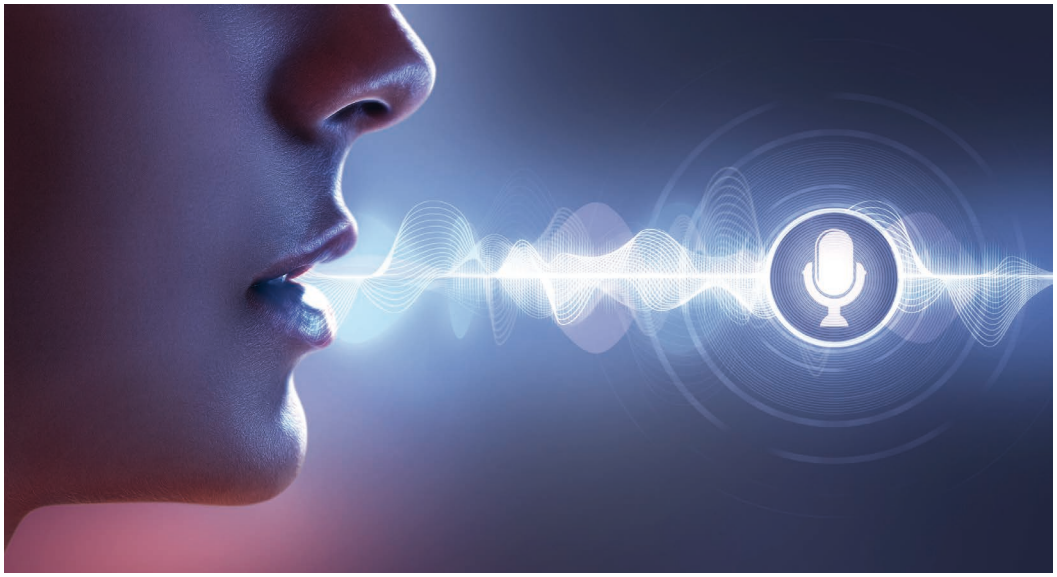
Die Verwendung von KI auf der Hardwareplattform des Gerätes kann schwierig sein: Umfangreiche Schaltungen mit hoher Leistungsaufnahme und entsprechendem Preis sind nicht ökonomisch realisierbar. Stattdessen verwendet man die KI in der Trainingsphase des Sprachsystems. Das Ergebnis wird dann auf die Hardwareplattform übertragen, die dann nur noch als Execution Engine agiert und daher mit wenigen Ressourcen in Hardware und Software auskommt.

Motivationstreiber Corona

Die pandemische Situation hat die Tendenz befördert, nicht mehr jedes Bedienelement berühren zu wollen. Kann eine Aufgabe durch Sprachbedienung erledigt werden, ist dieser Kontakt überflüssig. Sind die Hände nicht frei, nicht sauber oder feucht, kann ebenfalls die Sprache herhalten. Möchte man auch noch „den Kopf freihaben“ und das Ergebnis nicht auf einem Display ablesen, hilft die Ausgabe in synthetischer Sprache. Die aktuelle Technologie geht weit über das hinaus, was in den 80er-Jahren auf Homecomputern unter „Sprachausgabe“ verstanden wurde. Prosodie (Sprachmelodie) und Phrasierung klingen sehr natürlich; Satzzeichen strukturieren den angesagten Text.

HY-LINE verfolgt mit der HMI 5.0-Strategie die Absicht, möglichst viele Sinne zur Interaktion zwischen Mensch und Maschine einzusetzen – dort, wo es sinnvoll ist. So steht die Partnerschaft zu Voice Inter Connect aus Dresden unter dem Vorzeichen, das gesprochene Wort in die Kommunikation einzube-

Autor:
Rudolf Sosnowsky,
Leiter Technik
HY-LINE Computer Components
Vertriebs GmbH
www.hy-line.de



© *AdobeStock 231199039*

ziehen, sei es als Eingabemedium zur Steuerung der Maschine oder als Ausgabe für deren Status. Eine wichtige Rolle spielt dabei auch das GUI, das eingegebene Befehle und deren Auswirkungen für den Anwender aufbereitet darstellt.

Spracheingabe mit Natural Language Understanding

Die Ansprüche an eine bestimmte Technologie sind im professionellen Einsatz ungleich höher als im Smart Home-Umfeld. Die nahe 100 % liegende Verfügbarkeit und Zuverlässigkeit spielen hier eine eminente Rolle. Ist es im Smart Home nur eine Unannehmlichkeit, wenn das Licht mal nicht auf Kommando eingeschaltet wird („Entschuldigung, das Nachtlicht ist gerade nicht erreichbar!"), so ist es im professionellen Einsatz undenkbar, die OP-Leuchte nicht neu zu fokussieren oder den Braten im Dampfgarer nicht abzuschalten.

Eine Analyse zeigt, dass bei Systemen, die an eine Cloud angebunden sind, Latenzen auftreten, die zu hoch sind. Offline-Systeme sind hier klar im Vorteil: nicht nur arbeitet das System deterministisch und in Echtzeit, auch bleiben die Daten lokal und damit privat. Ohne den Zwang zu einer Anbindung an eine leistungsfähige Cloud, in der die Anfragen ausgewertet und bearbeitet werden, funktioniert das Gerät auch dort, wo eine Internet-Abdeckung fehlt, Daten nur mit einer mäßigen Bandbreite übertragen werden oder der Cloudanbieter seinen Service einstellt.

Hybrides Konzept

Das hier vorgestellte Konzept arbeitet hybrid: Das rechenintensive Training, bei dem die Sprachmodelle erstellt werden, findet auf einem leistungsfähigen Server in der Cloud statt. Nur das Ergebnis wandert in den lokalen Speicher und wird im Betrieb zur Erkennung der Eingabe verwendet. Dadurch reicht dem lokalen Rechner ein moderater Durchsatz aus, was sich in Wärmeentwicklung und Leistungsaufnahme positiv niederschlägt. Das bedeutet, dass die Sprachbedienung in der Ausführung rein auf dem lokalen System läuft und ohne Anbindung zur Laufzeit auskommt.

Bild 1: Starterkit für Sprachbedienung, © HY-LINE

Sprachausgabe per Text to Speech

Sprachsynthese macht aus der Sprachsteuerung mit Fokus auf Spracheingabe ein voll umfängliches Assistenzsystem mit Sprachausgabe auch für umfangreiche Texte. So kann sich der Bediener oder Servicetechniker aus einer hinterlegten Bedienungsanleitung mit Hilfe der passenden Suchbegriffe die relevanten Textpassagen heraussuchen und vorlesen lassen. Während der Fehlerbehebung bleiben die Augen weiter auf die Maschine gerichtet.

Auch hier hilft die KI bei der Erstellung der Synthesemodelle mit Machine Learning-Algorithmen, um bei der Text to Speech-Ausgabe Fließtexte in eine dynamische, natürlich klingende Sprachausgabe umzuwandeln. Wie bei dem Training der Spracherkennung ist der Prozess hier ebenso zweistufig: Training in der Cloud, Interpretation und Wiedergabe nur lokal – damit bleiben Daten vertraulich und sicher.

Argumente für Sprachbedienung

Warum ist die Bedienung mit Sprache so interessant und wichtig? Sie ist einfach zu verstehen und intuitiv zu nutzen. Nach dem Wake Word, mit dem das System aufgeweckt und zum Zuhören aufgefordert wird, können in natürlicher Sprache Befehle gegeben oder Informationen abgerufen werden. Im Idealfall ist es möglich, das System als „Do What I Mean"-Maschine zu nutzen. Ein Argument für die Bedienung ist auch, dass Sprache schneller kommuniziert als über ein anderes Eingabemedium wie z.B. die Tastatur. Der Weg im Hirn vom Gedanken zum Sprachzentrum ist kürzer als der Umweg, die Fingermuskeln in der richtigen Reihenfolge anzusteuern und damit eine Tastatur zu bedienen.

Vielfältige Anwendungsbereiche

Hauptmedium ist immer noch die manuelle Eingabe, ob mit Tastatur, Maus, Gestensteuerung oder ganz einfach über Bedientaster. Sprache kann die Eingabe überall dort ersetzen, wo die Hände nicht zur Verfügung stehen, weil sie anderweitig verwendet werden oder schmutzig sind. Dazu gehört das HMI an der Maschine in der Fertigungslinie, wo beide Hände für das Werkstück gebraucht werden, oder das Informationssystem am Point of Sale, das Auskunft erteilt, wo Läden in der Einkaufspassage oder Produkte in den Regalen zu finden sind.

In der Gastronomie kann beim professionellen Küchengerät die Temperatur auf das Grad genau eingestellt werden, während die Hände für Lebensmittel sauber bleiben. In der Logistik gibt das Lagersystem Anweisungen, wo ein Artikel entnommen oder abgelegt werden soll. In der Medizintechnik kommt es darauf an, die Hände steril zu halten und nicht zu verunreinigen, damit Viren und Bakterien nicht weitergetragen werden.

Auch neue Felder wie das Smart Caravaning sind für die Sprachbedienung geeignet: Wo heute Einzellösungen für das Schalten von Licht oder die Abfrage der Füllung von Frisch- oder Brauchwassertank

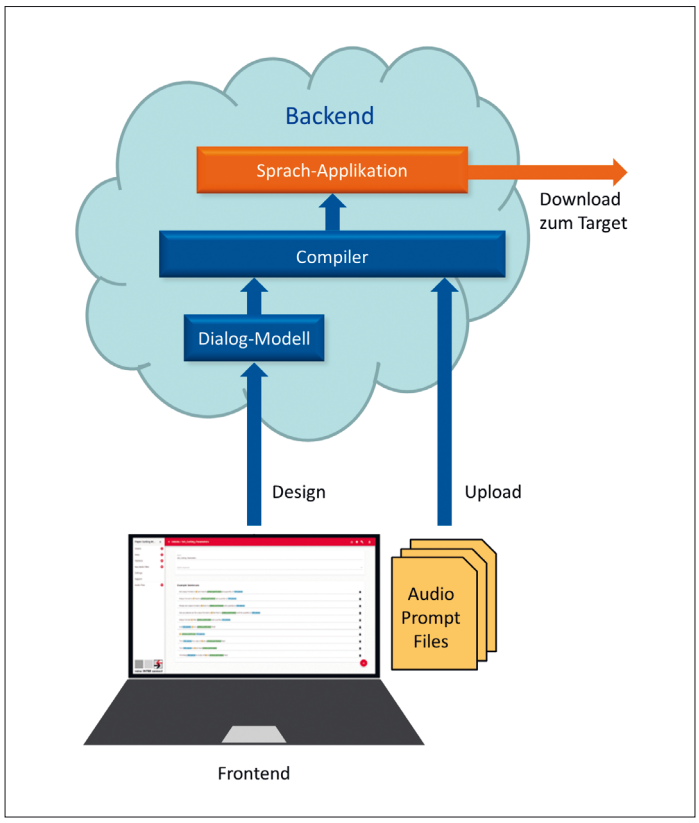


Bild 2: Entwicklung eines Sprachdialogs, © HY-LINE

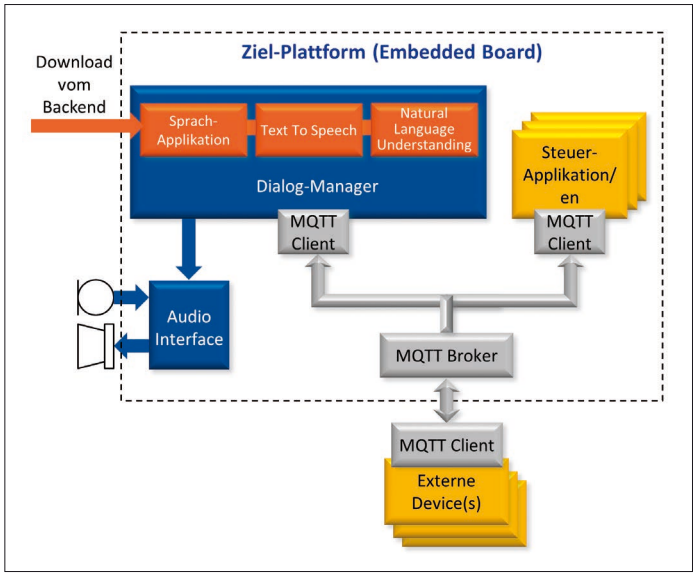


Bild 3: Ablauf zur Laufzeit, © HY-LINE

eingesetzt werden, kann eine einheitliche Oberfläche mit Sprachbedienung für eine einfachere Verdrahtung und ergonomischere Bedienung sorgen.

Kickstart zur professionellen Sprachbedienung

Mit einer fertigen Hardware- und Software-Lösung ist der Weg von der Idee bis zur fertigen Umsetzung einer Sprachbedienung nicht so steinig. Bild 1 zeigt das Starterkit, das nicht nur die ersten Schritte einfacher macht. Um ein Gerät zu entwickeln, das professionellen Ansprüchen genügt und rund um die Uhr im Einsatz ist, steht ein Web-SDK zur Verfügung, das die erforderlichen Algorithmen und Modelle abstrahiert. Unterschiedliche Sprachen sind bereits in Modulen hinterlegt. Der Entwickler erstellt das SUI (Speech User Interface) für die individuelle Anwendung mit spezifischen Dialogen und Befehlen. Darunter liegt das Maschineninterface, das Befehle des SUI an Hardware und GUI weitergibt.

Um diesen Prozess bis zur individuellen Sprachanwendung möglichst einfach zu gestalten, hat HY-LINE das Starter-Kit entwickelt, das nicht nur die ersten Schritte auf dem Weg zu einer kommerziellen Lösung einfacher macht.

Die Software

Als Teil des Starter-Kits steht ein Web-SDK zur Verfügung, mit dem die Beispiele weiter erkundet und eigene Applikationen erstellt werden können. Ganz ohne Programmierung werden eigene Dialogmodelle erstellt, indem Bedienphrasen mit Schlüsselwörtern eingegeben und auf dem Server kompiliert werden. Das Ergebnis wird dann auf das Starter-Kit heruntergeladen und funktioniert ohne Internet-Anbindung.

Iterativ wächst das Sprachsystem, indem Synonyme als Alternativ-Eingaben und weitere Befehlssätze formuliert werden. Die Architektur nimmt den Text entgegen und erkennt selbständig Schlüsselworte, die es als Subjekt oder Prädikat zuordnet. Füllworte wie „bitte“ und „äh“ werden übersprungen. Das SDK stellt APIs zur Verfügung, die über MQTT an das Gerät übergeben werden können. Damit wird der erkannte Sprachbefehl in eine Hardwareaktion umgesetzt. Dieser sind keine Grenzen gesetzt; die Reaktion kann in einer Sprachausgabe, dem Schalten eines Ports, einer Ausgabe auf dem Display oder der Änderung eines Wertes in einem JSON-File liegen. Das Kit ist vielseitig genug, um externe Geräte anzusteuern, so dass mit ihm funktionsfähige Prototypen erstellt und die Akzeptanz in der Zielgruppe getestet werden kann.

Die Hardware

Angetrieben wird das Sprachbedienungs-Kit von einem Single-Board-Computer im picoITX-Format, der auf der leistungsstarken iMX8.M-CPU basiert. Das Bedieninterface ist ein 10,1-Zoll-Display mit HD-Auflösung und kapazitivem Touchscreen. Alle Komponenten sind für den industriellen Einsatz geeignet, so dass eine kommerzielle Umsetzung auch mit dem Starter-Kit erfolgen kann. Natürlich kann die so erstellte Applikation auch auf eine andere Zielplattform portiert werden. Dies spart Zeit und Kosten bei der individuellen Sprachanwendung.

Die akustische Ausgabe kann im einfachsten Fall mit einem Summer erfolgen. Besser wird allerdings ein Lautsprecher eingesetzt, der breitbandig Quittierungstöne und Sprachmeldungen ausgeben kann. Während frühere Systeme zuvor aufgenommene Audio-Schnipsel zusammensetzten, um Meldungen auszugeben – wie etwa bei der Ansage von Uhrzeit und Datum – bietet mittlerweile Text to speech (TTS) die Freiheit, beliebige Texte in beliebigen Sprachen aus einem Textfile auszugeben. Der Wortschatz ist damit praktisch nicht limitiert und funktioniert genau wie die Spracheingabe lokal auf dem System ohne Internetverbindung zur Laufzeit.

Ablauf einer Implementierung

Mit Hilfe einer webbasierten Entwicklungsumgebung sind die folgenden Schritte erforderlich, um ein System für die eigene Anwendung zu definieren. Der Sprachdialog, also das Aktivierungswort, mit dem die Aufmerksamkeit des Systems auf Eingabe hergestellt wird, die zulässigen Kommandos und deren Parameter, werden im Webtool als Texteingabe zusammengestellt (Bild 2). Während der Eingabe findet bereits der erste Verarbeitungsschritt statt: Grapheme, also eingegebene Zeichen, werden in Phoneme, also kleinste akustische Bestandteile der Sprache umgewandelt.

Sind alle Worte definiert, werden mit den KI-basierten Algorithmen die definierten Sprachressourcen in ein statistisches und ein semantisches Modell übersetzt und zum Download angeboten. Das Ergebnis wird auf die Zielplattform heruntergeladen und gestartet. Dann kann der Netzwerkstecker gezogen werden – das Endprodukt läuft autark. Der Ablauf in der fertigen Applikation ist in Bild 3 dargestellt.

Audio-Technologie

Erstaunlich sind die Fähigkeiten des Gehirns, mit zwei Ohren und der Geometrie des Kopfes Geräusche zu isolieren und andere ganz aus-

Kickstart für professionelle Anwendungen

Touchless: Berührungslose und hygienische Bedienung

Industry Grade: Hohe Zuverlässigkeit und Verfügbarkeit, Echtzeitfähig

Do what I mean: Natürlich-sprachliche Kommunikation

Privacy by Design: Hohe Datensicherheit durch lokale Ausführung

Zero Coding: Einfachste webbasierte Sprachdialogerstellung

Text to Speech: Echter Dialog mit Sprachsynthese

zublenden. So gelingt es uns, auch an einem Tisch im Restaurant mit vielen Gästen uns auf das Gespräch mit dem Gegenüber zu fokussieren, die ebenso redenden Nachbarn und das Geklapper des Geschirrs aber auszublenden. Für ein Sprachsystem ist dies nicht so einfach. Erst durch die Hilfe eines Richtmikrophons oder elektronischer Filter erzielt das System eine ebenso hohe Erkennungsqualität durch Steigerung des Signal-Stör-Abstands. Das Richtmikrofon muss dabei nicht die lange Bauform haben, die man aus TV-Interviews kennt. Ein Array (Anordnung) mehrerer Einzelmikrofone erlaubt, auch aus einer lauten Umgebung den Sprecher des „Wake Words“ zu identifizieren und ihm bei Bedarf zu folgen. Damit steigert sich die Erkennungsgenauigkeit, die Reaktionsgeschwindigkeit und die Akzeptanz des Systems enorm. Die gleiche Technologie lässt sich auf der Audio-Ausgabeseite verwenden, um den Schall gezielt in eine Richtung abzustrahlen.

Fazit

Mit der Ergänzung durch Sprache gewinnt jedes User Interface eine neue Dimension. Die Implementierung ist einfacher als gedacht, denn mit dem Starterkit kann nicht nur sofort ein Demo gestartet, sondern auch erste Schritte mit eigenen Kommandos und Ausgaben gegangen werden. Für die Implementierung von Protokollen zur Ansteuerung externer Geräte steht ein leistungsfähiges SDK zur Verfügung. Durch die State-of-the-art-Technologie arbeitet das System unabhängig vom Sprecher; 30 Sprachen sind vordefiniert. Auch auf Plattformen mit beschränkten CPU- und Speicher-Ressourcen kann diese Lösung eingesetzt werden. Unter Umständen reicht hier auch ein digitaler Signalprozessor. Know-How in der Verarbeitung und Aufbereitung von Audio-Signalen garantiert ein zuverlässiges, schnelles System ganz ohne Online-Verbindung.

Wer schreibt

HY-LINE Computer Components steht als Mitglied der HY-LINE-Gruppe mit 30 Jahren Expertise als Spezialist für komplette Systemlösungen im Bereich Display- und Touchtechnologie und Embedded Computing auf Chip- und Boardebene. Sie bietet ihren Kunden neben der Distribution auch die Entwicklung von anwendungsspezifischen Produkten für Lösungen in den Bereichen Wireless, IoT, Leistungselektronik, Stromversorgung und Energiespeicher an. Als Leader in Technology bietet HY-LINE tiefe technische Beratung sowohl in State-of-the-Art-Technologien als auch innovativen Ansätzen. Ein Schwerpunkt bildet HMI 5.0, die Schnittstelle zwischen Mensch und Maschine, die multi-sensuelle und multi-modale Kommunikation bietet.

Weitere Informationen

https://www.hy-line-group.com/sprachsteuerung
www.voiceinterconnect.de ◄

Beispiele aus der Praxis

Bedienung und Überwachung von Medizintechnik

- Betten und Untersuchungsliegen für CT/MRT via Sprache steuern (z. B. Herauf- und Herunterfahren, Sitz- und Liegeposition)
- Informationsabfrage aus einer Datenbank, z. B. während einer Operation
- Dokumentation von Tätigkeiten in Kranken- und Altenpflege oder Rehabilitation
- Berührungsfreie Bedienung von Geräten in schlecht zugänglichen oder sterilen Umgebungen
- Unterstützung von Menschen mit Handicap, z. B. Sehschwäche oder Einschränkung der Bewegungsfreiheit

Industrie

- Natürliche Kollaboration mit Robotern und in Augmented Reality-Anwendungen
- Steuerung von Maschinen und Geräten
- Industrieautomation, Test- und Messtechnik
- Berührungsfreie Bedienung von Geräten in schlecht zugänglichen oder explosionsgefährdeten Umgebungen
- Qualitätssicherung, z. B. Abhaken einer Checkliste in beliebiger Reihenfolge

Point of Sales

- Info-Stelen in Einkaufszentren oder Servicepunkten
- Bestell-Automaten im Fast-Food-Restaurant
- Berührungslose Bedienung

Glossar

API	Application Program Interface: Software, die die grundlegende Kommunikation zwischen Anwendungsprogramm und der Hardware oder anderen Softwaremodulen zur Verfügung stellt
GUI	Graphical User Interface: Designerische Gestaltung der Bedienelemente auf einem Display zur Erzielung einer guten UX
HMI	Human Machine Interface: Computergestützte Schnittstelle zwischen Mensch und Maschine. Meist als Touchscreen-Terminal für Ein- und Ausgabe ausgeführt
JSON	JavaScript Object Notation: Datenaustauschformat, das für Menschen einfach zu lesen und zu schreiben und für Maschinen einfach zu parsen (Analysieren von Datenstrukturen) und zu generieren ist
MQTT	Message Queuing Telemetry Transport: Offenes Netzwerkprotokoll für die Kommunikation von Maschinen untereinander
SDK	Software Development Kit: Software-Umgebung für eine beschleunigte Entwicklung eines Anwendungsprogramms
SUI	Speech User Interface: Kommunikation über Spracheingabe und -ausgabe
TTS	Text to Speech: Beispielsweise in ein Terminalfenster eingegebener Text wird als Audio (über den Lautsprecher) ausgegeben
UX	User eXperience: Beschreibt, wie einfach (ergonomisch, physisch) ein Gerät zu bedienen ist