Hallo Werkzeugmaschine!

von Rudolf Sosnowsky

Sprachsteuerung ist in unserem Alltag angekommen, sei es im Auto, zur Bedienung des Smartphones oder des Heimsystems. Auch im professionellen Umfeld gewinnt die Technologie zunehmend an Bedeutung. Deren Einsatz ist dabei weniger kompliziert als gedacht.

Im Fokus
Human Machine Interfaces

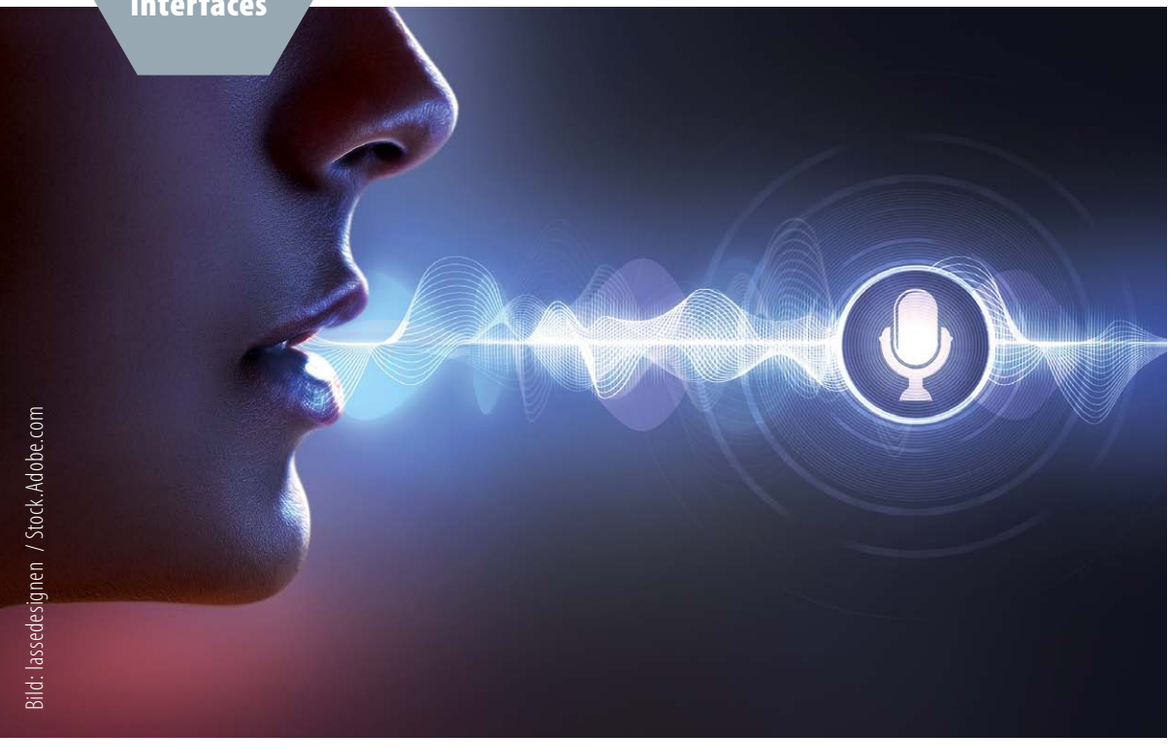

Bild: lassedesignen / Stock.Adobe.com

Zuerst als nette Spielerei betrachtet, dann im Smart Home zu einem Bestandteil der Einrichtung gereift: Die Steuerung von Musik, Licht, Erinnerungs-Timern und das Befüllen von Einkaufslisten ist mit dem Medium Sprache einfach und bequem. Während die Sprachbedienung anfänglich 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

Neben dem traditionellen Display- und Touchscreen-Interface steht dem Medium Sprache mit dem gesprochenen Wort als Eingabekommando und der synthetisierten Sprache als Ausgabe ein fester Platz als Bedienelement zu. Das Consulting-Unternehmen Gartner erstellt Studien für die Zukunft von Technologien. Der so genannte „Gartner Hype Cycle" erklärt dabei die Lebensphasen einer Technologie in mehreren Stufen – von der anfänglichen Euphorie über die Ernüchterung bei der Realisierung bis hin zum produktiven Einsatz.

Die Spracherkennung hat bereits die Phase der Produktivität erreicht, auf einem guten Wege dazu ist die Sprachsynthese. In das Verstehen und Interpretieren natürlicher Sprache ist aber noch Entwicklungsarbeit zu legen. Eine hohe Bedeutung nimmt neben der rein algorithmische die durch Künstliche Intelligenz (KI) unterstützte Spracherkennung ein. Doch was brauchen wir für den Einsatz im professionellen Umfeld? Was unterscheidet diese Anwendungen von gängigen Sprachassistenten?

Im Sinne eines ergonomisch designten HMI erwartet man von einer Spracherkennung eine sprecherunabhängige Erkennung des gesprochenen Wortes, das Verstehen möglichst mehrerer Sprachen, genau hin- und auch wegzuhören (manchmal wird die Sprachbedienung getriggert, wird das Schlüsselwort fälschlich erkannt), und tolerant bezüglich der Grammatik zu sein. Füllwörter wie „bitte", „einmal", „ja, genau" und Räuspern sollen ignoriert werden und nicht zu Fehlbedienungen führen.

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

Warum Sprachbedienung?

Die pandemische Situation hat die Tendenz befördert, nicht mehr jedes Bedienelement berühren zu wollen; sind die Hände zudem nicht frei, nicht sauber oder feucht, kann eine Aufgabe durch Sprachbedienung erledigt werden. Möchte man das Ergebnis nicht auf einem Display ablesen, hilft die Ausgabe

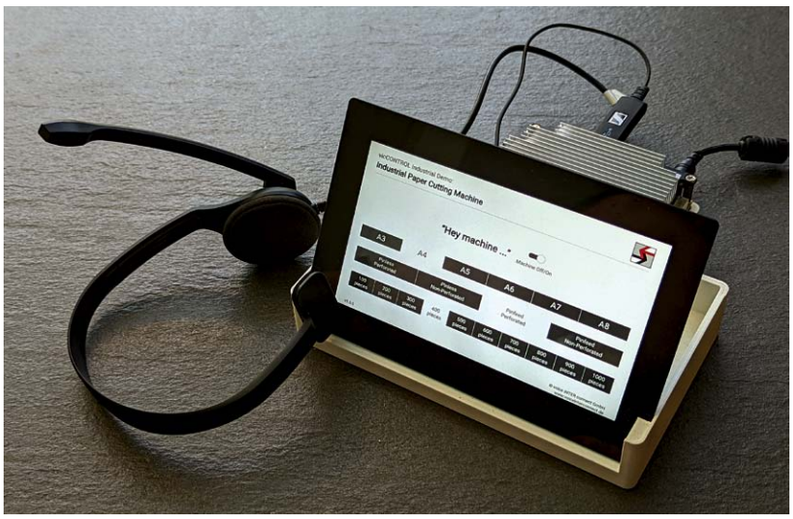

Bild1. Starter-Kit für Sprachbedienung. (Bild: Hy-Line)

in synthetischer Sprache. Die aktuelle Technologie ist mit der „Sprachausgabe" auf Heim-Computern aus den achtziger Jahren nicht mehr zu vergleichen. Prosodie (Sprachmelodie) und Phrasierung klingen sehr natürlich, Satzzeichen strukturieren den angesagten Text.

Warum ist die Sprachbedienung 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, dass Sprache schneller kommuniziert als über ein anderes Eingabemedium wie eine Tastatur. Der Weg vom Gedanken zum Sprachzentrum ist kürzer als der Umweg, die Fingermuskeln anzusteuern und damit eine Tastatur zu bedienen.

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, Dresden, unter dem Vorzeichen, das gesprochene Wort in die Kommunikation einzubeziehen, 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 Technologie sind im professionellen Einsatz ungleich höher als der im Smart Home-Umfeld. Die nahe 100 % liegende Verfügbarkeit und Zuverlässigkeit spielen hier eine eminente Rolle. Ist es im Smart Home eine Unannehmlichkeit, wenn das Licht nicht auf Kommando einschaltet, ist es im professionellen Einsatz undenkbar, die OP-Leuchte nicht neu zu fokussieren. Eine Analyse zeigt, dass bei Systemen, die an eine Cloud angebunden sind, Latenzen auftreten, die zu hoch sind. Offline-Systeme sind klar im Vorteil: nicht nur arbeitet das System deterministisch und in Echtzeit, die Daten bleiben 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 die Abdeckung durch das Internet nicht vorhanden ist, Daten nur mit einer mäßigen Bandbreite übertragen werden oder der Cloudanbieter seinen Service einstellt.

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 nieder-

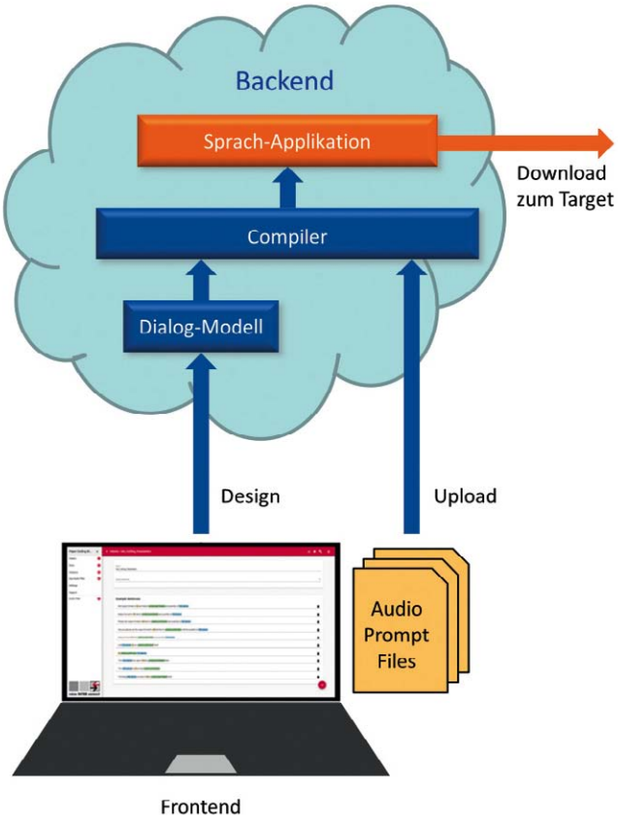

Bild 2. Entwicklung eines Sprachdialogs. (Bild: Hy-Line)

schlägt. Das bedeutet: Die fertige Applikation zur Sprachbedienung läuft rein auf dem lokalen System, ohne eine Anbindung an einen Server über das Internet zu benötigen.

Sprachausgabe mit „Text to Speech"

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

Bei der Erstellung der Synthesemodelle für „Text to Speech" (TTS) kommt KI zum Einsatz: Die Modelle nutzen Machine Learning-Algorithmen, welche dabei helfen, Fließtexte in eine dynamische, natürlich klingende Sprachausgabe umzuwandeln. Wie bei dem Training der Spracherkennung ist der Prozess auch hier zweistufig: Training in der Cloud, Interpretation und Wiedergabe nur lokal – damit bleiben Daten vertraulich und sicher.

Kickstart für die professionelle Sprachbedienung

Hauptmedium ist immer noch die manuelle Eingabe – ob mit Tastatur, Maus, Gestensteuerung oder 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 zählen beispielsweise das HMI an der Maschine in der Fertigungslinie, wo beide Hände für das Werkstück gebraucht werden, oder das Informationssystem am Point of Sales, welches 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 das Lebensmittel sauber bleiben. In der Logistik gibt das Lagersystem Anweisungen, wo ein Artikel entnommen oder abgelegt werden soll. In der Medizintechnik kommt darauf an, die Hände steril zu halten oder nicht zu verunreinigen, damit Viren und Bakterien nicht weitergetragen werden. Auch neue Felder wie Smart Caravaning sind für die Sprachbedienung geeignet: Wo heute Einzellösungen für das Schalten von Licht oder die Abfrage der Füllung von Frisch- oder Brauchwassertank eingesetzt werden, kann eine einheitliche Oberfläche mit Sprachbedienung für eine einfachere Verdrahtung und ergonomischere Bedienung sorgen.

Eine fertige Hardware- und Software-Lösung ebnet den Weg von der Idee bis zur fertigen Umsetzung einer Sprachbedienung. **Bild 1** zeigt das Starter-Kit, 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 WebSDK zur Verfügung, das die erforderlichen Algorithmen und Modelle abstrahiert. Unterschiedliche Sprachen sind bereits in Modulen hinterlegt. Der Entwickler erstellt das SUI 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 ein 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 erkundet und eigene Applikationen erstellt werden können. Ohne Programmierung werden eigene Dialogmodelle erstellt, indem Bedienphrasen mit Schlüsselwörtern eingegeben und auf dem Server kompiliert werden. Das Ergebnis wird 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 Hardware-Aktion umgesetzt. Diese Reaktion kann in einer Sprachausgabe, einem 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 iMX8.M-CPU basiert. Das Bedieninterface ist ein 10,1-Zoll-Display mit HD-Auflösung mit kapazitivem Touchscreen. Alle Komponenten sind für den industriellen Einsatz geeignet, so dass eine kommerzielle Umsetzung mit dem Starter-Kit erfolgen kann. Die so erstellte Applikation kann auch auf eine andere Zielplattform portiert werden. Die akustische Ausgabe kann im einfachsten Fall mit einem Summer erfolgen. Besser wird

Web-Tipp

Ein Glossar mit den gängigen Abkürzungen sowie eine Liste verschiedener Anwendungsbeispiele für Sprachbedienung in der Praxis finden Sie unter
https://bit.ly/3JoW6fC

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 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.

Die Audio-Technologie

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

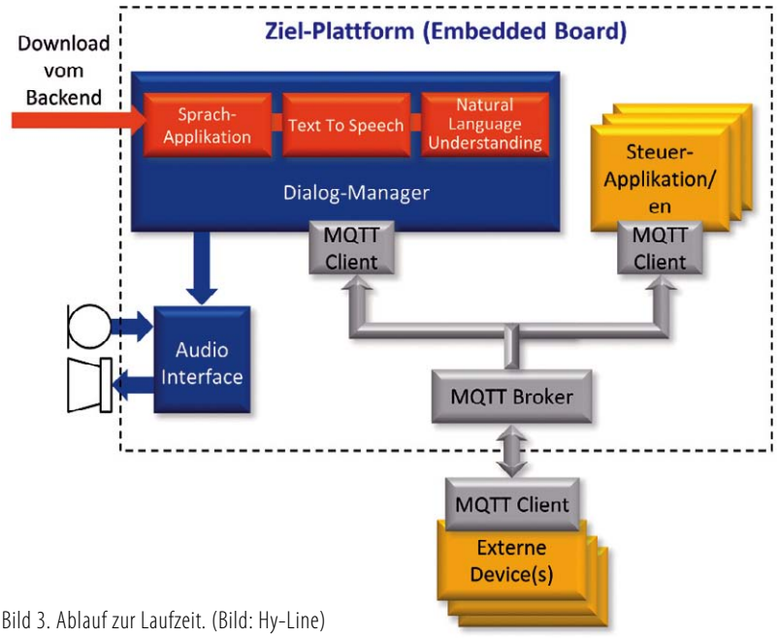

Bild 3. Ablauf zur Laufzeit. (Bild: Hy-Line)

Neue Dimensionen erschließen

Mit der Ergänzung durch Sprache gewinnt jedes User Interface eine neue Dimension. Die Implementierung ist einfacher als gedacht, denn mit dem Starter-Kit kann nicht nur eine Demo gestartet werden, 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. ag

Rudolf Sosnowsky
ist Leiter Technik bei Hy-Line Computer Components Vertriebs GmbH in Unterhaching.

Im Interview

Welche Vorteile hat Sprachbedienung? Wie steht es um die Akzeptanz der Technologie in der Industrie? Eignet sie sich auch auf Saftey-Applikationen? Martin Dibold von Hy-Line gibt im Interview weitere Einblicke: **bit.ly/3JlW8Vi**