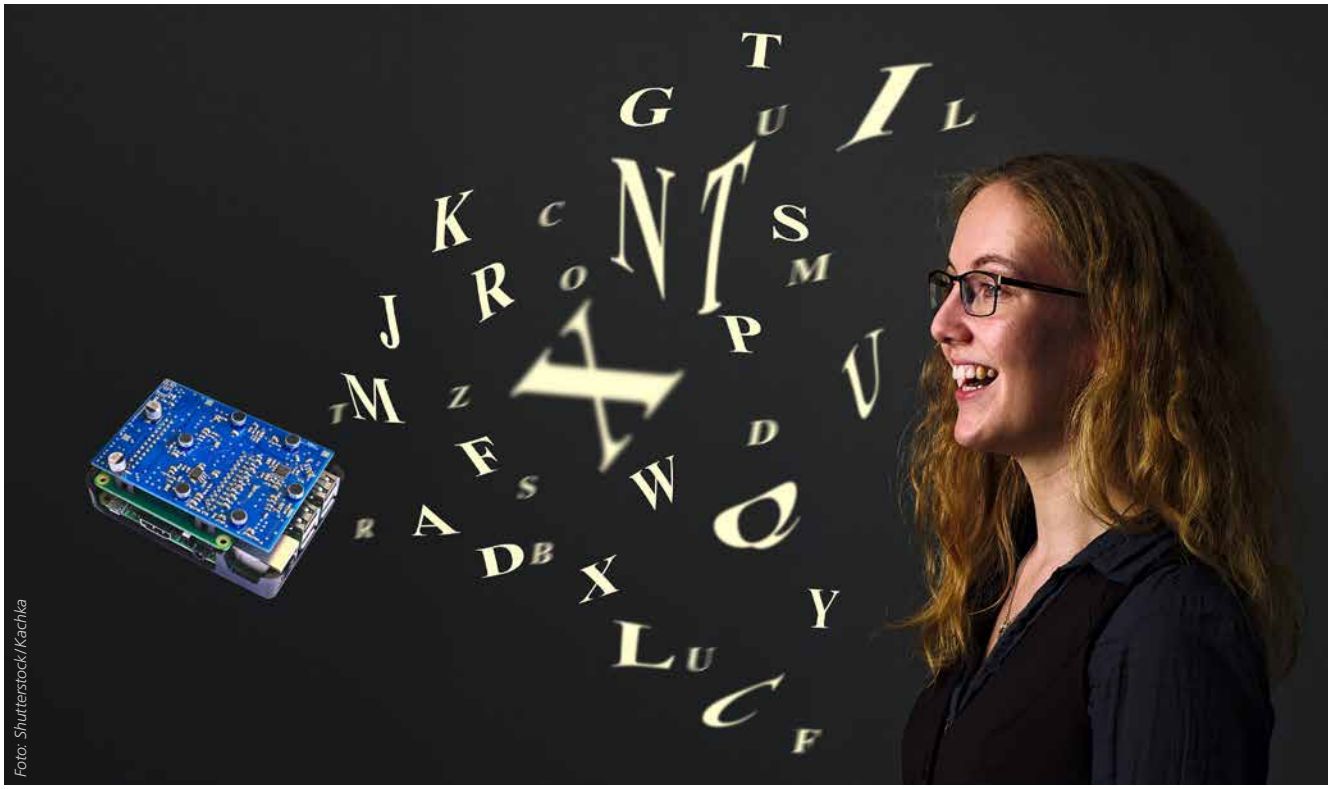
Reinhold Häb-Umbach



Lektionen für Alexa & Co?!

Wenn akustische Signalverarbeitung mit automatisiertem Lernen verknüpft wird: Nachrichtentechniker arbeiten mit mehreren Mikrofonen und tiefen neuronalen Netzen an besserer Spracherkennung unter widrigsten Bedingungen. Von solchen Sensornetzwerken könnten langfristig auch digitale Sprachassistenten profitieren.

Apples „Siri“, Amazons „Alexa“ oder andere smarte Alltagshelfer und -begleiter sind inzwischen fast allgegenwärtig. Digitale Assistenten, die über Sprachbefehle bedient werden, haben in den letzten Jahren einen nicht für möglich gehaltenen Siegeszug angetreten. Sie reagieren bekanntlich „aufs Wort“ und spielen zum Beispiel die vorher angewählte Musik ab, führen Einkaufslisten oder beantworten Fragen.

Mittlerweile gibt es Zehntausende sogenannter Skills, die Aufgaben übernehmen können – sei es nun das Hochdrehen der Heizung oder das Herunterlassen der Fensterjalousien im „smart home“. Voraussetzung ist neben dem Sprachbefehl nur, dass das Gerät oder die Anwendung an das Internet oder an das Heimnetzwerk angebunden sind.

Grundlagen- und anwendungsorientierte Forschung sowie neue Technologien haben dafür die Wege bereitet. Die Forschung an Spracherkennungssystemen, also an Computerprogrammen, die menschliche Sprache in eine maschinenlesbare Darstellung umsetzen können, begann in den 1960er-Jahren. Vieles blieb aber damals erfolglos. Die entwickelten Systeme konnten zwar einige Dutzend Einzelwörtern erkennen, allerdings nur unter Labor-

Oben: Was der Mensch kommuniziert, wird als „Sprachsignal-Wolke“ von einer leistungsfähigen Mikrofonanlage aufgezeichnet. Rechts: ein „akustischer Sensorknoten“, bestehend aus einer mehrkanaligen Mikrofongruppe und einem Minirechner mit Funkmodul.

bedingungen, das heißt in ungestörten Umgebungen. Dies lag einerseits an dem begrenzten Wissen in diesem neuen Forschungsgebiet, aber andererseits auch an den begrenzten technischen Möglichkeiten vor 50 Jahren.

Dann ein Sprung: Ab Mitte der 1980er-Jahre fanden Methoden der Wahrscheinlichkeitsrechnung Einzug in die automatische Spracherkennung und verbesserten die Erkennungsleistung maßgeblich. Die ersten kommerziellen Systeme, die Fließtext mit praktisch unbegrenztem Vokabular erkennen konnten, wurden in den 1990er-Jahren auf den Markt gebracht. Es waren Spezialsysteme für dedizierte Anwendungen im professionellen Umfeld, etwa zur Verschriftlichung von medizinischen Befunden. Denn nur wenn die Anwendung klar umrissen war, konnten die Systeme eine einigermaßen zufriedenstellende Sprachverarbeitung erreichen.

Es sollten weitere 20 Jahre bis zu einem neuen Durchbruch vergehen. Zu Beginn der 2010er-Jahre fanden die „tiefen neuronalen Netze“ Einzug in die Sprachverarbeitung. Mit dieser als „deep learning“ bezeichneten Technik konnte die Erkennungsleistung deutlich gesteigert werden. So berichteten beispielsweise Forscher von Microsoft und IBM im Jahr 2017, dass sie „human parity“ erreicht hätten: Der Computer könne gesprochene Sprache mit der gleichen Wortfehlerrate erkennen wie ein Mensch.

Dabei ist das Konzept der künstlichen neuronalen Netze zunächst denkbar einfach, und die eingesetzten Lernalgorithmen sind seit Mitte der 1980er-Jahre bekannt. Das Netz besteht im Wesentlichen aus der Hintereinanderschaltung von Schichten, in denen jeweils eine Multiplikation von Eingangszahlenwerten mit aus Trainingsdaten gelernten Zahlen, den sogenannten Gewichten, erfolgt und die Resultate aufsummiert werden. Auf das Ergebnis wird anschließend eine Nichtlinearität angewendet und die resultierende Größe an die nächste Schicht weitergereicht. Bei den heute verwendeten tiefen Netzen kann die Anzahl der Schichten groß sein und inzwischen mehrere Hundert betragen.

Beim Lernen, dem „Training“ des Netzes, geht es darum, die Gewichte aus Trainingsdaten zu bestimmen. Sie bestehen aus Sprachaufnahmen samt den zugehörigen Texten. Legt man das erste Element an den Eingang und das zweite als Trainingsziel an den Ausgang des Netzes, werden die Gewichte so bestimmt, dass sie für ein gegebenes Audiosignal am Eingang den Text am Ausgang möglichst gut vorhersagen. Anschließend kann das Netz dann beliebige Spracheingaben transkribieren und in Text umsetzen.

Wenn die grundlegenden Algorithmen schon in den 1980er-Jahren bekannt waren, wieso kam es erst so viel später zu diesem Durchbruch? Wesentliche Erfolgsfaktoren waren die eminent höhere Leistungsfähigkeit heutiger Computer und die verfügbaren, durchaus gigantischen Sprachdatenmengen. Sie umfassen mehr als 1000 Stunden gesprochener Sprache. Erst auf dieser Grundlage

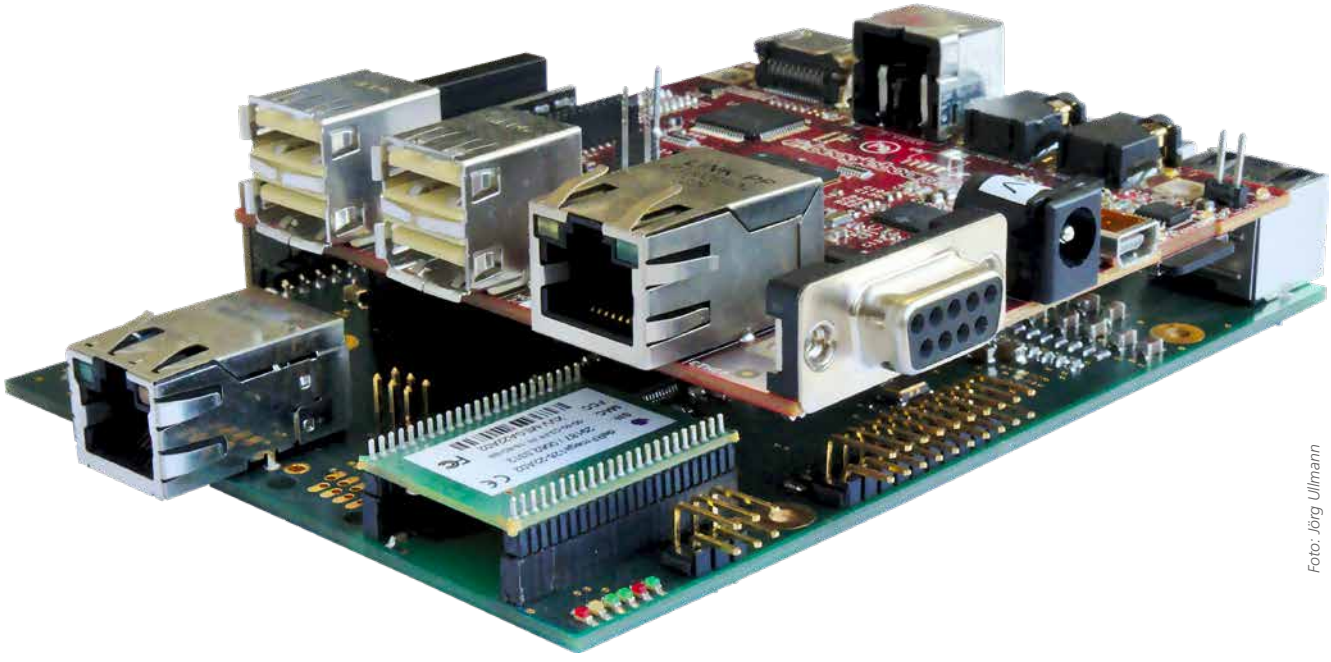


Foto: Jörg Ullmann

wurde das Trainieren tiefer Netze technisch möglich.

Der Durchbruch bei tiefen neuronalen Netzen motivierte auch zu Anwendungen im Konsumbereich. Hierbei war der digitale Assistent „Siri“ auf den Apple iPhones ein Vorreiter. Bis zu „Alexa“ oder ähnlichen Geräten, die im Wohnzimmer stehen und aus der Ferne bedient werden, war es jedoch ein langer Weg. Denn bei den komplexen Anwendungen war die gewünschte Spracherkennung aus zwei Gründen ungleich schwieriger: Zum einen kann das Sprachsignal durch Raumhall und andere Signalquellen im Raum (wie Nebengeräusche des Fernsehens) gestört werden, zum anderen muss die Bedienung vollständig über Sprache erfolgen. Eine Bedienung mithilfe Tastatur oder berührungsempfindlichem Bildschirm ist nicht mehr möglich.

Ein wesentlicher Erfolgsfaktor war der Einsatz von Mikrofongruppen statt nur eines Mikrofons. Damit lässt sich eine „Richtkeule“ auf die Sprachsignale der Zielperson ausrichten, wobei Signale aus anderen Raumrichtungen unterdrückt werden. Eine solche Ausnutzung räumlicher Information ist mit einem einzelnen Mikrofon nicht möglich.

Zuverlässige Spracherkennung bei entfernten Mikrofonen ist ein Forschungsthema an der Universität Paderborn. Gefördert mit Mitteln der DFG wurden klassische Methoden der „Sensorarraytechnik“ mit neuronalen Netzen verknüpft, um die Spracherkennung und -verarbeitung auch in stark gestörten Umgebungen zu verbessern. Den Erfolg dieses Ansatzes belegen internationale Vergleichstests und die Tatsache, dass führende Forschergruppen weltweit mittlerweile den Ansatz übernommen haben und nun ihrerseits weiterentwickeln.

Vielleicht kann ein einprägsames Bild helfen, den regulierenden Effekt der Signalverarbeitung auf die Mustererkennung mit neuronalen Netzen zu erläutern: Ein neuronales Netz zur Spracherkennung benötigt Trainingsdaten aus repräsentativen Erkennungssituationen. Vergleichbar mit einem Schrotgewehr wird das System mit Sprachdaten gefüttert, respektive Schrotkugeln. Das geschieht in der Hoffnung, dass in einer Anwendung die konkrete Erkennungssituation dabei ist, das heißt, dass bei einem Schuss auf das Ziel eine der Schrotkugeln wohl treffen wird. Mithilfe der Signalverarbeitung macht man das Gewehr zielsicherer, man benötigt nach und nach nicht mehr so viele Schrotkugeln, um das Ziel zu treffen.

Die geschickte Verknüpfung von klassischer Signalverarbeitung und neuronalen Netzen birgt Vorteile gegenüber einer Lösung, die rein auf neuronale Netze setzt. Sie er-

Laboraufbau in der Nachrichtentechnik, mit dessen Hilfe eine neu entwickelte mehrkanalige Mikrofongruppe ausgetestet wird.

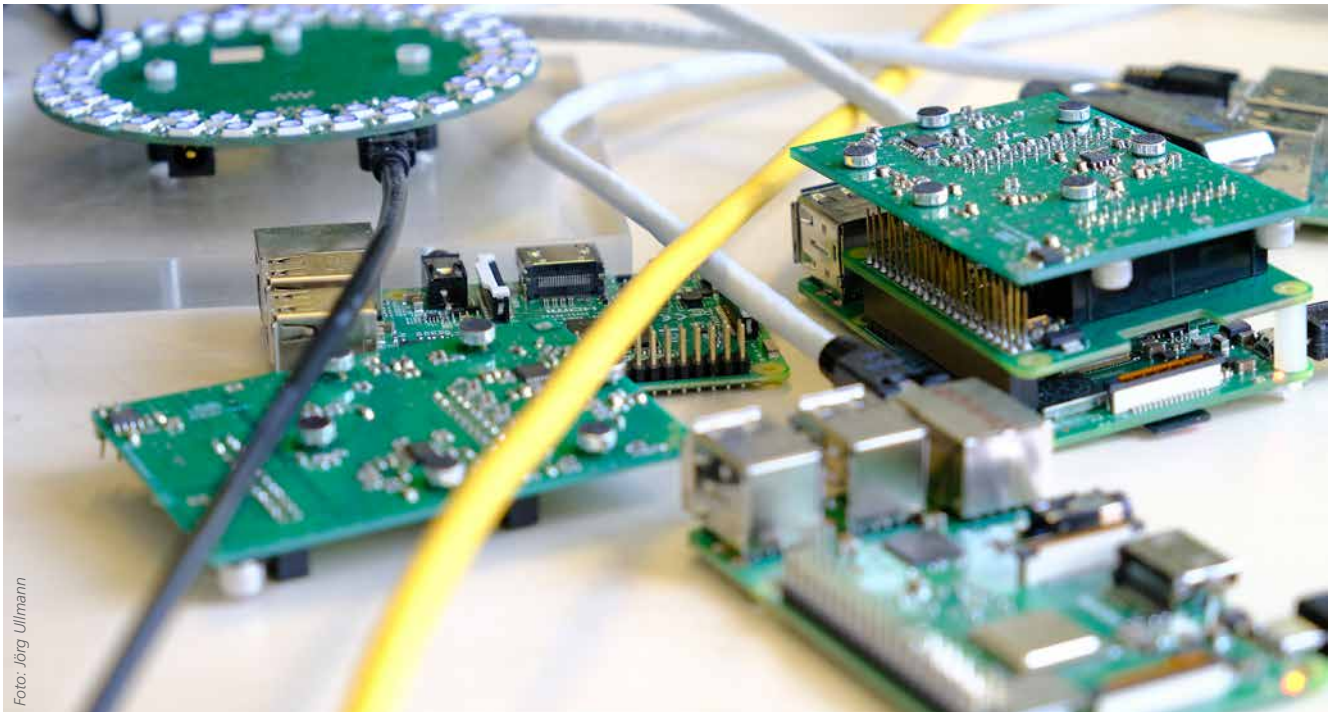


Foto: Jörg Ullmann

laubt es, Vorwissen über eine Problemstellung einzubringen, benötigt weniger Trainingsdaten und Rechenleistung und erlaubt schließlich eine bessere Interpretation der Verarbeitungsschritte im Vergleich zu der „black box“-Lösung eines neuronalen Netzes. Wenn es um die Verbesserung der Sprachverarbeitung geht, etwa die Enthallung eines Signals oder die Trennung eines Gemischs mehrerer Sprecher, zeigt sich dieser hybride Ansatz als erfolgreich.

Worin liegt die Zukunft der digitalen Assistenten? Mittlerweile ist es nicht unüblich, in einem Haushalt mehr als einen intelligenten Lautsprecher anzutreffen. Da jedes Smartphone mit mindestens einem Mikrofon ausgestattet ist und auch viele andere Geräte mittlerweile über ein Mikrofon verfügen (zum Beispiel die Fernbedienung des Fernsehers), liegt die Frage nahe, ob mit diesen verteilten Mikrofonen nicht eine viel bessere Signalerfassung möglich ist als mit dem einen Gerät, das irgendwo im Raum steht. Die Wahrscheinlichkeit ist groß, dass sich eines der Mikrofone in der Nähe des Sprechers befindet.

Solche Szenarien sind im Blick der DFG-Forschungsgruppe „Akustische Sensornetzwerke“, die Kommunikations-, Signalverarbeitungs- und Mustererkennungsaspekte von Netzwerken verteilter Mikrofone erforscht. Andere großflächigere Anwendungen im außerhäuslichen Bereich sind beispielsweise die akustische Überwachung von Lärmschutzvorgaben oder von Artenschutzreservaten. Eine wichtige Rolle spielen dabei auch Überlegungen zum Schutz der Privatsphäre. Ein Stichwort ist hierbei „Privacy by Design“: Bereits bei der Signalerfassung wird, wiederum mithilfe

Grafik: Shutterstock / Kachka

Wie kann ein Smartphone noch leistungsfähiger gemacht werden? Die Verbesserung von Sprachaufnahmen auch bei Störgeräuschen ist zum Beispiel ein Ansatzpunkt.

neuronaler Netze, das Signal so stark komprimiert, dass nur noch die Zielanwendung realisierbar ist (zum Beispiel die Erkennung von Vogelstimmen).

Sind die digitalen Assistenten, die auch gerne als „intelligente Lautsprecher“ bezeichnet werden, denn nun intelligent? Mitnichten, wie das Gedankenexperiment des chinesischen Zimmers des Philosophen John Searle aufzeigt. Hierbei stellt man sich einen geschlossenen Raum vor, in dem ein Mensch, der keinerlei Chinesisch versteht, in chinesischer Schrift gestellte Fragen anhand einer in seiner Muttersprache verfassten Anleitung in chinesischer Schrift sinnvoll beantwortet: Durch einen Schlitz wird ihm ein Zettel mit einer in Chinesisch verfassten Frage hereingereicht. In einer großen Tabelle sucht er die ihm dargereichten Zeichen und findet als Tabelleneintrag diejenigen Zeichen, die er als Antwort herausreichen soll. Personen außerhalb des Raums folgern aus den Ergebnissen, dass der Mensch in dem Raum Chinesisch beherrscht, obwohl das nicht der Fall ist. Algorithmen oder kompetente Geräte, allein und für sich betrachtet, sind nicht intelligent, aber sie können „intelligent“ eingesetzt und genutzt werden.

Prof. Dr. Reinhold Häb-Umbach
ist Inhaber des Lehrstuhls für Nachrichtentechnik an der Universität Paderborn und Sprecher der DFG-Forschungsgruppe „Akustische Sensornetzwerke“.

Adresse: Universität Paderborn, Fakultät für Elektrotechnik, Informatik und Mathematik (EIM), Institut für Elektrotechnik und Informationstechnik, Warburger Straße 100, 33098 Paderborn

DFG-Förderung in der Einzelförderung und im Rahmen der Forschungsgruppe „Akustische Sensornetzwerke“.

www.uni-paderborn.de/asn