Unterrichtseinheit: Spracherkennung

Thema

Aufbau von Computersystemen und Speichern von Daten in Kombination mit ausgewählten Aspekten des Lernfeldes Daten und ihre Spuren insbesondere die rechtlichen Rahmenbedingungen für den Umgang mit ihren persönlichen Daten wie z.B. informationelle Selbstbestimmung, Allgemeine Geschäftsbedingungen (AGB) und Datenschutz.

Kurzbeschreibung

	**Spracherkennung (Kurzbeschreibung)**
Idee	In dieser Unterrichtssequenz werden sowohl die technischen Vorgänge bei der Spracherkennung als auch Datenschutzfragen in Zusammenhang mit der zentralen Erfassung von Sprachprofilen problematisiert
Organisatorisches	Schulform: Sekundarstufe 1 Klassenstufe: 7-10 Zeitumfang: ca. 40 Minuten

	**Spracherkennung (Kurzbeschreibung)**
Curriculares Umfeld	**Kerncurriculum Informatik Sek 1 Niedersachsen** Primäre Lernfelder: „Daten und ihre Spuren“ & „Automatisierte Prozesse“ - Die Schüler:innen unterscheiden zwischen lokalen und verteilten Anwendungen. - Die Schüler:innen unterscheiden zwischen Informationen und ihrer Repräsentation durch Daten. - Die Schüler:innen diskutieren ausgewählte Fälle in Bezug auf die Grundlagen des Datenschutzes. - Die Schüler:innen beschreiben die einzelnen Schritte beim Ablauf eines automatisierten Prozesses. - Die Schüler:innen nennen gesellschaftliche Konsequenzen des Einsatzes automatisierter Prozesse, z. B. in der industriellen Produktion. **Orientierungsrahmen Medienbildung** - Die Schüler:innen entwickeln ein Bewusstsein für Datensicherheit, Datenschutz und Datenmissbrauch, um ihre Privatsphäre durch geeignete Maßnahmen zu schützen. (4.2) - Die Schüler:innen beurteilen verwendete digitale Werkzeuge im Hinblick auf den Datenschutz und mögliche gesellschaftliche Auswirkungen. (5.3) - Die Schüler:innen finden Lösungen für technische Probleme und verstehen Funktionsweisen sowie grundlegende Prinzipien der digitalen Welt. (5.3) - Die Schüler:innen schätzen die Auswirkungen digitaler Technologien auf wirtschaftliche, soziale und politische Prozesse ein. (6.3)
Lernziele	- Indem die Schüler:innen Audiodaten codieren und eine visuelle Repräsentation der resultierenden Daten sehen, erhalten sie eine Vorstellung von der Verarbeitung von Sprache in IT-Systemen - Dadurch dass Schüler:innen unterschiedlich eindeutige visuelle Repräsentationen von Audiodaten decodieren, wird ein Bewusstsein für den komplexen technischen Prozess der Umwandlung von Speech2Text geschaffen. - Durch ein visuelles Modell des Datenflusses abstrahieren Schüler:innen den Weg von Daten, der sich exemplarisch auf weitere Prozesse übertragen lässt (Geodaten, Trackinginformationen) - Indem Schüler:innen die Probleme der zentrale Verarbeitung personenbezogene Daten in Cloudsystemen diskutieren, wird ihnen die gesellschaftliche Dimension bewusst.
Voraussetzungen	Lehrer:innen und Schüler:innen sind mit der Bedienung eines Webbrowsers vertraut. Sie können Screenshots anfertigen und im Dateisystem ablegen. Der Lehrkraft ist das Definitionsproblem rund um die Termini „künstliche Intelligenz“ und „adaptives System“ bewusst. Die Begriffe Information, Daten und Codierung sind sicher in der Lerngruppe eingeführt.

	**Spracherkennung (Kurzbeschreibung)**
Ablauf	Die Schüler:innen bekommen in einem Erklärvideo das Phänomen Spracherkennung („Speech to Text“) anhand der Diktierfunktion auf dem iPad vorgeführt. Sie stellen Vermutungen über die technischen Prozesse dahinter an. Die technischen Herausforderungen werden anhand von Aufzeichnungen phonetischer Laute als Hüllkurve von den Schüler:innen erarbeitet. Auf Basis der Ergebnisse kann sowohl die technische Notwendigkeit der Datenverarbeitung in Cloudsystemen als auch die Datenschutzproblematik thematisiert werden. Eine innere Differenzierung ist durch Zusatzaufgaben wie etwa das Anfertigen eines Diagramms zum schematischen Ablauf der Spracherkennung möglich.
fachlicher Hintergrund	Um gesprochene Sprache in Text umwandelt zu können, muss diese zunächst vom Endgerät in Audiodaten umgewandelt werden. Diese werden cloudbasiert analysiert und das Ergebnis wird dann an das Endgerät zurückübermittelt. Bei der Analyse der Rohdaten treten zahlreiche Herausforderungen auf, etwa sehr ähnliche, anhand von reinen Audiodaten kaum unterscheidbare Laute. Daher kommen sowohl Mechanismen der adaptiven Datenverarbeitung (Verbesserung der Erkennungsrate durch mehr Daten) als auch phonetische und grammatische Strategien zur Anwendung (z.B. Buchstabenergänzung bei bestimmten Lautkombinationen).
Materialien	Onlinewerkzeug für Sprachaufnahmen

Spracherkennung als Phänomen im Alltag

Spracherkennung in reiner Form etabliert sich zunehmend auf Streamingplattformen wie YouTube oder Twitch. Videos lassen sich in Echtzeit untertiteln bzw. werden das zunehmend automatisiert. Sprachassistenten unterstützen Nutzer:innen im einfachsten Fall bei der Eingabe von Daten in IT-Systeme, etwa bei Navigationssystemen. Hier werden die gewonnenen Daten bereits durch IT-Systeme interpretiert und in Geodaten umgesetzt. Die Interpretation von Sprachdaten ist bei Sprachassistenten wie Alexa, Siri oder Cortana noch offensichtlicher. Die Grenzen der Spracherkennung wird momentan vor allem in Telefonanlagen mit Sprachmenus spürbar. Alexa fordert explizit zur Verwendung von Sprachprofilen auf, was einen Hinweis darauf gibt, dass bei der Spracherkennung auch die personalisierte Datenerhebung eine Rolle spielt.

Gesellschaftliche Relevanz

Spracherkennung ermöglicht die vereinfachte Eingabe komplexer Texte in IT-Systeme. Dabei berücksichtigt das jeweilige IT-System Grundsätze der Grammatik und Rechtschreibung. Um die Qualität der Spracherkennung zu steigern, werden vermutlich viele Daten in Rechenzentren zusammengeführt und verarbeitet. Die zugrundeliegenden Algorithmen unterliegen i.d.R. keine gesellschaftlichen Kontrolle. Die Stimme eines Menschen ist einzigartig und ein personenbezogenes

Merkmal. Der Anwendungszweck „Spracherkennung“ ist technisch ohne die Verarbeitung personenbezogener Datenanteile möglich und wird im Datenschutzrecht als Zweckbestimmung ausdrücklich gefordert.

Spracherkennung kann bei der Strafverfolgung dazu dienen, große Datenmengen - wie sie bei Audiodaten vorliegen - drastisch zu reduzieren und damit der automatisierten Verarbeitung überhaupt erst zugänglich zu machen. Dies kann aufgrund des erreichbaren hohen Automatisierungsgrades gesellschaftliche Fragen aufwerfen, wie sie im Kontext der Vorratsdatenspeicherung immer wieder diskutiert werden.

Die automatisierte Erkennung von Sprache ermöglicht neue Produkte im wirtschaftlichen Bereich, etwa die Entwicklung von Sprachassistenten. Damit werden unterschiedliche gesellschaftliche Bereiche adressiert.

Die Problematiken rund um den Prozess der Spracherkennung sind prototypisch für alle IT-Prozesse, die zu einer großen Akkumulation von Daten führen.

Analyse des Sachgegenstands

Spracherkennung basiert technisch auf einer Mustererkennung in komlexen Datenbeständen. IT-Systeme analysieren automatisiert digitalisierte Audioaufnahmen. Bei Vorliegen bestimmter typischer Merkmale kann mit einer gewissen Wahrscheinlichkeit auf eine phonetische Repräsentation durch Zeichen geschlossen werden. Dabei treten verschiedene Problemstellungen auf u.a.:

- Gleich klingende phonetische Laute lassen sich in Sprache durch unterschiedliche sprachliche Zeichen repräsentieren, z.B. bei Diphthongen (eu, äu / ai, ei)
- Das gleiche sprachliche Zeichen kann für mehrere phonetische Laute stehen (stimmloses und stimmhaftes „s“, „e“ im In- und Auslaut)
- die Stimmen von Sprecher:innen enthalten individuelle Merkmale (u.a. Oktavsprung bei Männern und Frauen)
- es gibt innerhalb einer Sprachgemeinschaft Varianten (Dialekte, Soziolekte etc.)
- bestimmte Lautgruppen lassen sich anhand von Audioaufnahmen nur sehr schwer unterscheiden (z.B. Nasale wie n und m)

Da diese Komplexität bewusst oder unbewusst von Menschen wahrgenommen wird, spricht man in der Alltagssprache im Kontext von Spracherkennung durch IT-Systeme gelegentlich von „künstlicher Intelligenz“. Die Definitionen von Intelligenz und künstlicher Intelligenz sind jedoch wissenschaft noch nicht abgesichert. Interessant ist z.B. die Liste der AGI Sentinel Initiative, die allein 17 Definitionen beider Begriffe gegenüberstellt[1)]. Als Arbeitsdefinition für das vorliegende Unterrichtsprojekt wird folgende Hilfsdefinition verwendet:

> *Künstliche Intelligenz bezeichnet die Fähigkeit von Computersystemen, auf sie zugeschnittene Aufgaben selbsttätig zu lösen, die aufgrund ihrer Komplexität bislang menschliche Fähigkeiten erforderten.*[2)]

Tatsächlich werden die skizzierten Probleme bei der Spracherkennung durch IT-Systeme durch

Kombination mit anderen Daten gelöst - so kommen schwer identifizierbare Laute in einer Sprache oft in bestimmten Buchstabenkombinationen vor und die konkrete Schreibung von Lauten folgt orthografischen Regeln. Der Sprachkorpus heutiger Sprachen „passt“ zudem problemlos in ein IT-System, sodass dieses sogar in Zweifelsfällen problemlos auf das Lexikon „zurückfallen“ „kann“.

IT-Systeme zur Spracherkennung sind dabei vor allem in ihrer Verarbeitungsgeschwindigkeit und ihrem Zugriff auf umfassende Datenbestände auf Spezialgebieten dem Menschen überlegen[3)]. Das wirkt auf informatisch nicht vorgebildete Personen beeindruckend. Trotzdem darf nicht verschwiegend werden, dass die allermeisten KI-Systeme meist nur eng definierte Aufgabenbereiche abdecken und in alternativen, z.B. psychologischen Definitionen nicht oder nur mit sehr großen Einschränkungen als „intelligent“ gelten können.

Die Strukturen hinter der Spracherkennung können im Rahmen dieses Unterrichtsprojektes teilweise dekonstruiert werden.

Da das Thema künstliche Intelligenz Wissenschaft noch extrem unscharf definiert ist, ergeben sich für eine Behandlung im Rahmen des Informatikunterrichts große Schwierigkeiten für eine schultaugliche Definition. Die einzig sinnvolle Möglichkeit ist letztlich die Problematisierung des Begriffes, die ein hohes Abstraktionsvermögen bei den Schüler:innen voraussetzt. Eine derartige Auseinandersetzung gehört damit in den Bereich der Differenzierung nach oben.

Kontext und Einbettung der Einheit in eine Mehrwochenplanung

Die Schüler:innen benötigen grundständiges Wissen zur Funktion des Internets und sollten den Begriff „Algorithmus“ in einem informatischen Sinne grob definieren können. Inhaltlich vertieft diese Einheit einen Teilaspekt von Sprachassistenten.

Daher bietet sich folgende Einbettung in ein didaktisches Umfeld an:

Material	**Inhalte (Kurzbeschreibung)**
Der Internetversteher (it2schhool Modul B2)	Funktionsweise des Internets verstehen
Finde die KI (it2schhool Modul KI-B1)	Künstliche Intelligenz im Alltag erkennen
Diese Unterrichtseinheit	Am Beispiel der Spracherkennung den Begriff KI sowohl problematisieren als auch „entzaubern“ durch Vertiefung eines Teilaspekts
Im Dialog mit KI (it2schhool Modul KI-B2)	Wiederaufnahme der Inhalte des Internetverstehers, Spracherkennung als Teilaspekt in einen größeren Zusammenhang stellen

Konkrete Unterrichtseinheit

Wie funktioniert die Diktierfunktion oder Spracherkennung von z.B. Alexa, Cortana, Siri etc.?

Schau dir dieses kurze Video an - am besten im Vollbildmodus (rechts unten in der Ecke des Videos aktvieren):

diktierfunktion_ipad.mp4

- Probiere das einmal aus mit einem deiner Texte
- Wie funktioniert das?
- Warum braucht das iPad dafür eine Internetverbindung?

... vielleicht finden wir gemeinsam eine Lösung.

Wie kann ein Computer Sprache erkennen?

Kurzeinführung in Audiomass

Erstmal muss man Sprache für den Computer „sichtbar“ machen. Das geht recht einfach mit dem Online-Audioeditor Audiomass. Du brauchst dafür keine App. Es reicht ein Browser wie Safari völlig aus.

Kurze Unterbrechung ...

Eigentlich möchtest du mit Sprache Informationen weitergeben, z.B. wie du dich fühlst oder was du gerne magst. Damit diese **Information** bei anderen ankommt, musst du sie aussprechen. Der Informatiker würde sagen: Du **codierst** die Information so, dass sie zu jemand anderem übertragen werden kann. Dann werden aus den Informationen **Daten**. Damit ein Computer Informationen aus deiner Sprache gewinnen kann, muss sie anders codiert werden. Genau das macht unser Online-Audioeditor.

Übrigens

Informationen, **Daten** und **Codierung** sind Fachbegriffe aus der Informatik. Jetzt aber weiter!

Wie du deine erste Aufnahme machen kannst, erkläre ich dir hier.

audiomass_intro.mp4

Mit Audiomass kannst du noch viel mehr machen, z.B. kleine Podcasts aufnehmen oder englische Texte oder Gedichtvorträge einsprechen. Deine Daten bleiben dabei auf deinem Gerät!

b und p - zwei einfache Buchstaben oder schon Laute?

Aufgabe 1

Mache Aufnahmen der Laute B und P. Du darfst dabei nicht so sprechen, als wenn du das ABC aufsagst. Du musst die Laute so aufnehmen, wie du sie sprichst, also wie das „B“ im Wort **B**aum oder das „P“ im Wort **P**latz.

Aufgabe 2

Zeige jetzt deine Aufnahmen jemandem aus deiner Klasse, der auch eine solche Aufnahme gemacht hat. Könnt ihr nur anhand des „Bildes“ unterscheiden, um welchen Laut es sich handelt?

m und n - Schon schwieriger?

Aufgabe 3

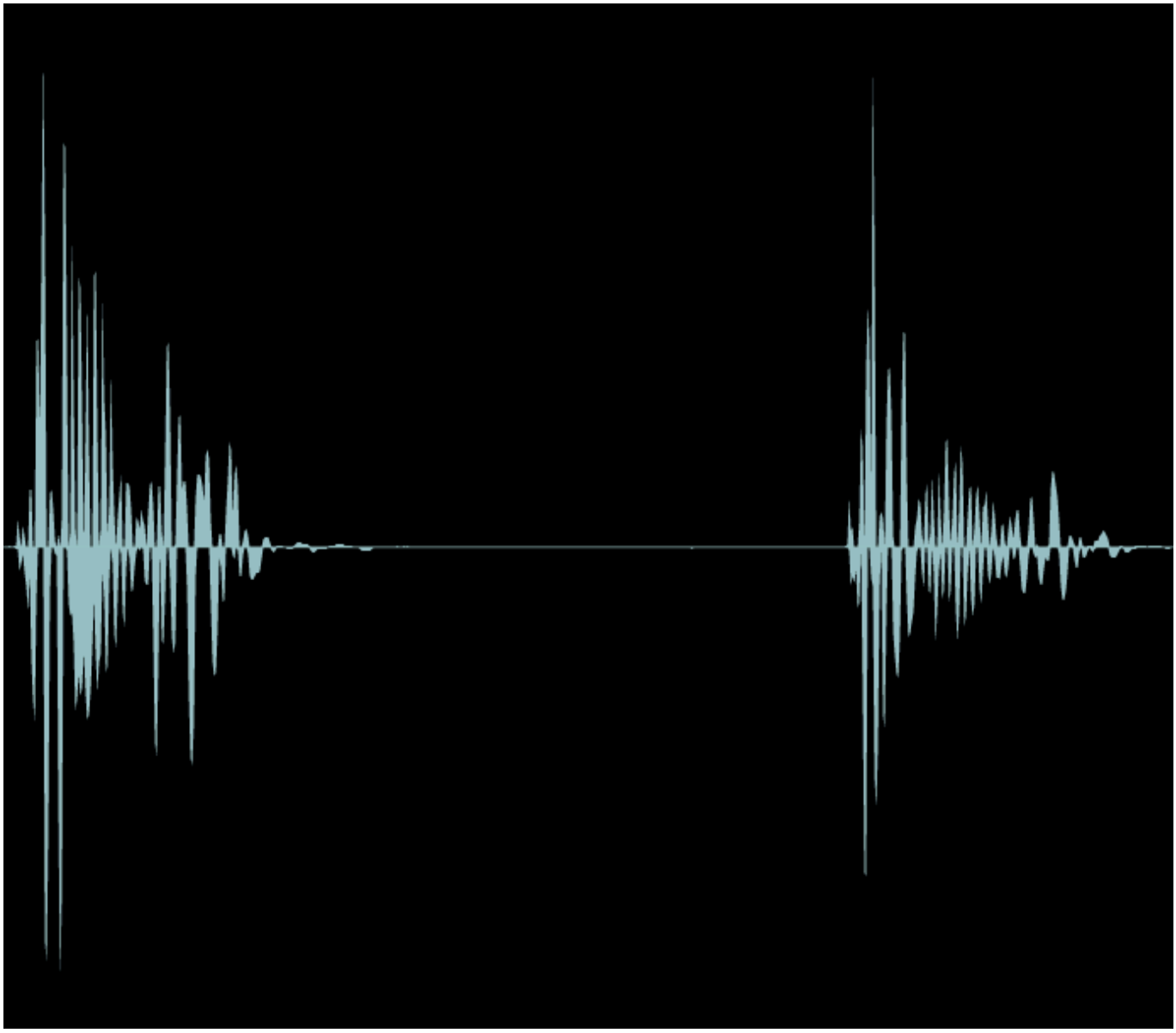
Mache Aufnahmen der Laute M und N. Du darfst dabei nicht so sprechen, als wenn du das ABC aufsagst. Du musst die Laute so aufnehmen, wie du sie sprichst, also wie das „N“ im Wort **N**ena oder das „M“ im Wort **M**oos.

Aufgabe 4

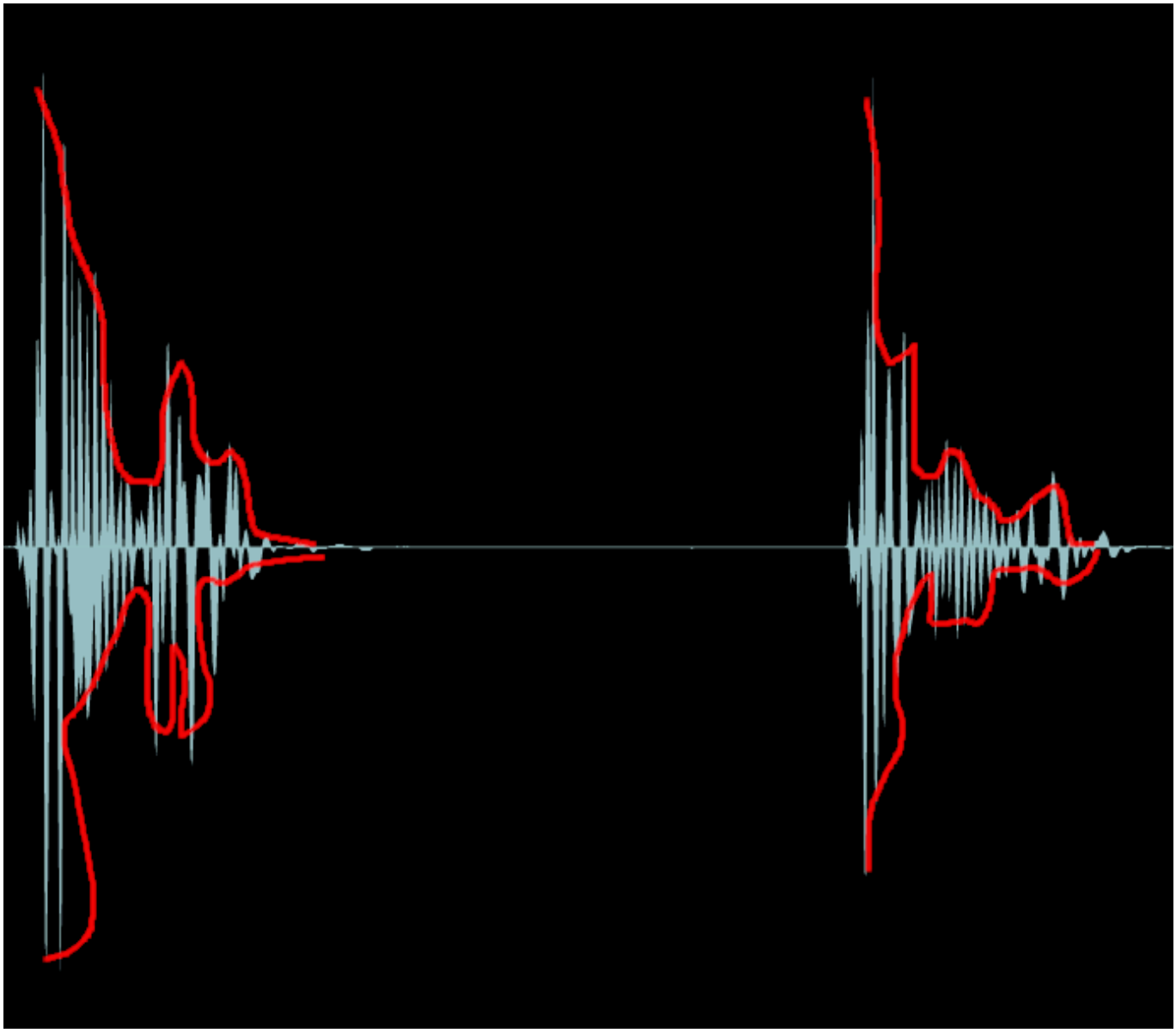
Wenn du Lust hast, kannst du das Ganze noch mit weiteren Lauten probieren, z.B. o und a oder k und t. Welche Laute kannst du gut erkennen/unterscheiden, welche nicht so gut?

Beispielergebnisse

Für B und P könnten die codierten Daten so aussehen (links: „B“, rechts: „P“):



Jetzt könnte man probieren, Muster zu finden, die typisch für einen Laut sind:



Anhand des Musters kann ein Computer entscheiden, um welchen Laut es sich handelt und dann ein „B“ oder ein „P“ ausschreiben. Beim „B“ gibt es hier offenbar eine größere „Lücke“ weiter hinten.

Und warum jetzt die Internetverbindung?

Tatsächlich ist rein technisch die Internetverbindung nicht mehr notwendig. Bei iPads funktioniert die Spracherkennung ganz ohne Internetverbindung. Bisher haben wir nur mit einzelnen Lauten herumexperimentiert - und das war schon gar nicht so einfach zu erkennen. Ganze Wörter bestehen aus vielen unterschiedlichen Lauten - da braucht es eine Menge Rechenleistung. Daher können zusätzlich die codierten Daten „über das Internet“ ein einen leistungsstarken Rechner gesendet werden, der dann aber nicht mehr die Umwandlung in Buchstaben vornimmt und an dein iPad zurückschickt, sondern diese Daten nutzt, um den Algorithmus zu optimieren, der auf dem iPad die Spracherkennung realisiert. Das geht aber so schnell, dass man fast keine Verzögerung merkt.

Was passiert da bei Apple und Co. wirklich in der Cloud?

So viel weiß man darüber tatsächlich nicht. Tatsächlich gibt es schon

Spracherkennungssysteme, die ganz ohne Cloud- oder Internetanbindung auskommen. Da ist es aber bei der Aktivierung nicht mit einem Erklärvideo getan. Was aber könnte in der Cloud tatsächlich geschehen?

- Daten könnten gesammelt werden, um die Spracherkennung immer besser zu machen
- Daten könnten gesammelt werden, um den Sprecher/die Sprecherin anhand seiner/ihrer Stimme zu erkennen
- Daten könnten gesammelt werden, um etwas über den Musikgeschmack oder die sprachlichen Kenntnisse von jemandem zu erfahren

Weil man das nicht so richtig weiß, gibt es durchaus Vorbehalte gegen den Einsatz von z.B. Sprachassistenten wie Siri, Alexa oder Cortana.

1) http://agisi.org/doc/AGISI_DefinitionsIntelligence.pdf

2) Gethmann, Buxmann Distelrath, Humm, Lingner, Nitsch, Schmidt, Spiecker genannt Döhmann: „Künstliche Intelligenz in der Forschung - Neue Möglichkeiten und Herausforderungen für die Wissenschaft“, S. 10, aus der Reihe: Ethics of Science and Technology Assessment Bd. 48, bei: Springer

3) Thai-Son Nguyen, Sebastian Stueker, Alex Waibel: Super-Human Performance in Online Low-latency Recognition of Conversational Speech. PrePrint: https://arxiv.org/abs/2010.03449