Hausarbeit im Modul „Aktuelle Themen“

Künstliche Intelligenz in der Audiowelt

Bence Böröcz (34898) – Kapitel 2
Marian Hepp (35424) – Kapitel 3
Maurice Schill (35425) – Kapitel 1
Hochschule der Medien Stuttgart
03.März 2019

Inhaltsverzeichnis

1. Künstliche Intelligenz in der Audiopostproduktion

1.1. Einleitung

Der Begriff der künstlichen Intelligenz umfasst einen großen Themenkomplex und ist aktueller denn je. Dahinter verbergen sich verschiedene Ansätze der Automatisierung und die Fragestellungen, wie Maschinen ein intelligentes, menschenähnliches Verhalten beigebracht werden kann. In dieser Ausarbeitung im Rahmen der Vorlesung „Aktuelle Themen: Künstliche Intelligenz und ihre Auswirkung auf Ihre Zukunft“, wird auf die grundlegende Idee und Funktionsweise von neuronalen Netzen und dem damit verbundenen Deep Learning eingegangen und wie diese in der Audiopostproduktion genutzt werden können. Dazu wird nachfolgend erläutert wie mit deren Hilfe Störgeräusche aus Audiodateien entfernt werden können. Diese Funktionsweise wird anhand der Software „Izotope RX 7 Advanced“ vorgestellt.

1.2. Neuronale Netze und Deep Learning

Neuronale Netze ähneln von der Funktionsweise und Struktur stark dem Gehirn von Menschen und Tieren. Diese Netze bestehen auch aus Einheiten, die Neuronen genannt werden. Die Neuronen sind alle miteinander vernetzt und in der Lage parallel zu arbeiten, in dem sie sich über gerichtete Verbindungen Informationen, in Form von Aktivierungssignalen, zusenden.[1] In Abbildung 1 ist eine schematische Darstellung eines neuronalen Netzes zu sehen. Hierbei sind die Eingabeneuronen rot, die versteckten Neuronen gelb und die Ausgabeneuronen grün dargestellt. Mithilfe der Eingabeneuronen werden Reize oder sonstige Informationen übergeben. Die versteckten Neuronen, auch Zwischenschicht genannt, befinden sich zwischen den Ein-

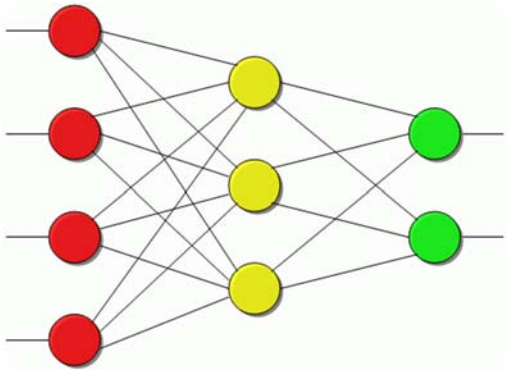

Abbildung 1: Schematische Darstellung eines einfachen neuronalen Netzes

[1] Vgl. Kruse et al. (2015), S.7

gabe- und Ausgabeneuronen.[2]

Das neuronale Netz ist tiefer, je mehr Zwischenschichten vorhanden sind. Die Anzahl dieser ist theoretisch unbegrenzt, diese benötigen pro hinzukommender Schicht mehr Rechenleistung.[3] Die Ausgabeneuronen übergeben die Signale bzw. das Ergebnis des Netzes.[4] Die neuronalen Netze sind die Grundvoraussetzung für das Deep Learning. Um komplexere Sachverhalte abbilden zu können, müssen mehrere Zwischenschichten und damit weitere Neuronen verwendet werden. Hier wird von Deep Learning gesprochen. Dies ist eine Methode der Informationsverarbeitung, um Prognosen und Entscheidungen treffen zu können. Große Datenmengen und ein definierter Prozess sind notwendig, um mithilfe des Deep Learnings Ergebnisse zu erhalten. Dieser Ablauf basiert auf der Funktionsweise des menschlichen Gehirns.[5]

1.3. Künstliche Intelligenz in der Software iZotope RX

Das „DeRustle-Modul“ in der Audioreparatursoftware iZotope RX Advanced wird verwendet, um Raschelgeräusche in Audioaufnahmen zu minimieren bzw. vollständig zu entfernen. Dieses Modul spielt vor allem in der Audiopostproduktion von Filmen eine große Rolle, da hier oft Ansteckmikrofone, auch Lavaliermikrofone genannt, verwendet werden, welche an die Kleidung des Schauspielers angebracht werden. Durch die Bewegung des Schauspielers entstehen Raschelgeräusche, welche sich häufig mit dem von ihm gesprochenen Dialog überlagern. Somit sind diese Audiodateien in der Audiopostproduktion oft nicht mehr verwendbar. Ein weiterer Faktor, der es erschwert Raschelgeräusche zu eliminieren, ist das Material der Kleidung, da jeder Stoff anders klingt. Außerdem handelt sich bei Rascheln um ein nicht statisches Signal. Wie in Abbildung 2 zu sehen, lässt sich dieses Störsignal nicht

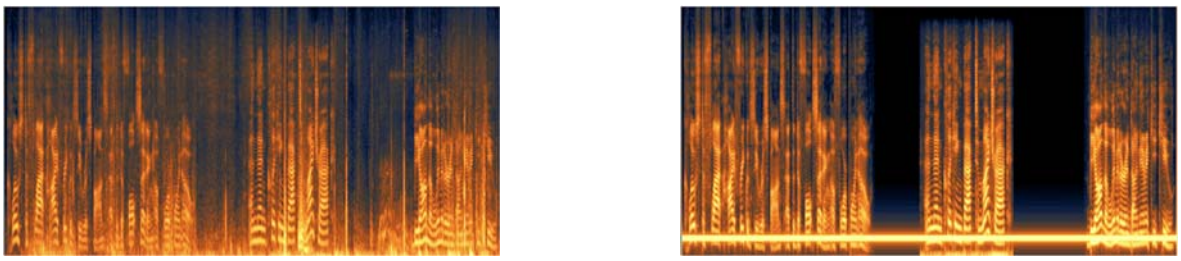

Abbildung 2: Verraschelter Dialog (links) und reiner Dialog mit 200 Hz Brummton (rechts)

[2] Vgl. Beck (2018).
[3] Vgl. Moeser (2017).
[4] Vgl. Beck (2018).
[5] Vgl. Litzel (2017).

sonderlich gut im Spektrogramm (Bildliche Darstellung eines Frequenzspektrums) einer Audiodatei identifizieren, wie es bei einem 200 Hz Brummton der Fall ist.

Daher lässt sich schlussfolgern, dass Rascheln mit den bekannten Signalverarbeitungsschritten wie Equalizern und Filtern nicht präzise genug minimiert werden kann. Die Firma Izotope Inc. setzt daher bei ihrem „DeRustle-Modul“ auf die Technik des Deep Learnings und der Quellentrennung.[6] In den nachfolgenden Unterkapiteln wird auf die beiden Aspekte eingegangen und erläutert die Vorgehensweise, die die Firma Izotope in ihrem DeRustle-Modul anwendet.

1.3.1. Quellentrennung in iZotope RX De-Rustle

Mithilfe der Quellentrennung wird versucht aus einer Einheit einzelne Elemente zu extrahieren. Es wird versucht, aus einer Sprachaufnahme mit Rascheln dieses und die „reine“ Sprache getrennt voneinander zu extrahieren. Dieses Verfahren kommt nicht nur in der Audiowelt zur Anwendung, sondern auch in den Bereichen der Neurowissenschaft oder in der Chemie. Im Audiobereich basiert dieses Phänomen auf dem „Cocktail Party Effekt“. Unter diesem Effekt wird die Fähigkeit des menschlichen Gehirns, bestimmte Audiosignale aus der Umgebung herauszufiltern oder sich auf bestimmte Audiosignale zu konzentrieren, verstanden. Das bekannteste Beispiel ist die Szenerie bei einer lauten Party, bei der versucht wird seinen Gesprächspartner zu verstehen. Daher stammt auch der Name „Cocktail Party Effekt“. Der nächste Schritt ist die Frequenzzuweisung, „spectral masking“ genannt. Hierbei wird für beide Quellen, Stimme und Rascheln, eine binäre Maske erstellt.

Mithilfe dieser Maskierung wird versucht den Frequenzbereich von 0 Hz bis 20 kHz einer der beiden Quellen zuzuordnen. Der Frequenzbereich wird durchlaufen und für jede Frequenz eine null oder eine eins in der jeweiligen Maske gespeichert. Somit kann jede Frequenz seiner Quelle zugeordnet werden, wobei einzelne Frequenzen ausschließlich zu einer Quelle (Maske) gehören können. Diese Maske kann als Verhältnismaske betrachtet werden, die Aufschluss darüber gibt, welcher Prozentsatz beider Quellen an jedem Zeitpunkt und jeder Frequenz vorhanden ist. Je höher der jeweilige Wert an einer Frequenz ist, desto mehr Anteil eines Signals/einer einzelnen Quelle ist in dieser Frequenz vorhanden.[7] In Abbildung 3 ist ein Frequenzspektrum

[6] Vgl. Izotope, Inc. (2017).
[7] Vgl. Izotope, Inc. (2017).

einer verraschelten Stimme zu sehen. Darunter befindet sich die vorhergesagte Verhältnismaske der reinen Stimme, die das Programm mithilfe der Quellentrennung bestimmt hat. Diese zeigt die angenommenen Frequenzen der Stimme ohne Rascheln. Um aus dieser Maske die reine Stimme zu erhalten, wird jeder einzelne Wert an der jeweiligen Frequenz der Verhältnismaske mit dem Frequenzspektrum der verraschelten Stimme multipliziert. Darauffolgend wird eine inverse Fouriertransformation durchgeführt, um das Audiosignal aus dem Frequenz- in den Zeitbereich zurück zu transformieren. Daraus resultiert eine Audiodatei mit der reinen Stimme.[8]

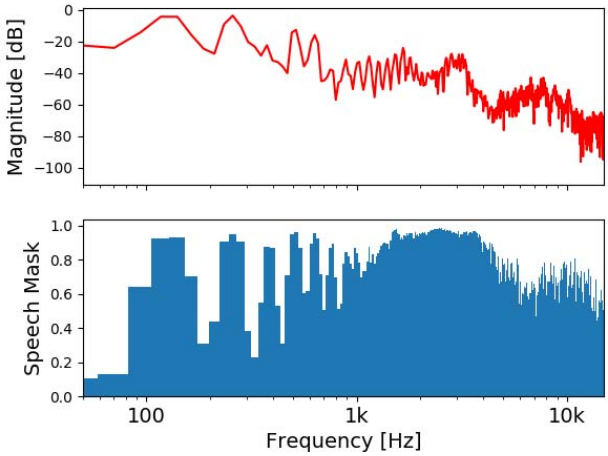

Abbildung 3: Frequenzspektrum einer verraschelten Stimme(oben) und die vorgesehene Verhältnismaske der reinen Stimme (unten)

1.3.2. Deep Learning in iZotope RX De-Rustle

Um die vorgestellte Quellentrennung anwenden zu können, bedarf es den neuronalen Netzen sowie dem Deep Learning. Diese helfen dabei die Aufnahme des Lavaliermikrofons in zwei Audiodateien zu trennen, eine mit der reinen Stimme und eine mit reinem Rascheln. Das Unternehmen iZotope verwendet das rekurrente neuronale Netzwerk.[9] Wie in Abbildung 4 zu sehen, bietet diese Form des neuronalen Netzes den Vorteil, dass Verbindungen zwischen der vorherigen und aktuellen Schicht möglich sind.[10] Wird dieser Vorteil auf die Problemstellung angewendet, kann jede im Netz stattgefundene Veränderung nachvollzogen und zugeordnet werden. Um als Resultat die reine Stimme zu erlangen, muss dieses Netz und das zugehörige Pro-

[8] Vgl. Izotope, Inc. (2017).
[9] Vgl. Izotope, Inc. (2017).
[10] Vgl. Haselhuhn (2018), S.4.

gramm trainiert werden. Hierfür wird dem neuronalen Netz zunächst eine Audiodatei übergeben, die ausschließlich reine Stimme enthält. Damit ist dem Netz das richtige Endergebnis bekannt. Im nächsten Schritt wird dieselbe Audiodatei mit Rascheln übergeben. Um mögliche Fehler beheben zu können, kann ein Fehlersignal ausgegeben werden, wodurch die Werte der einzelnen Neuronen angepasst werden können, die zu Abweichungen in Bezug auf die reine Stimme geführt haben. Dieser Vorgang wird Backpropagation genannt.[11] Aus diesen Daten werden die nötigen Informationen und Muster extrahiert, um das Netz Entscheidungen treffen zu lassen.[12]

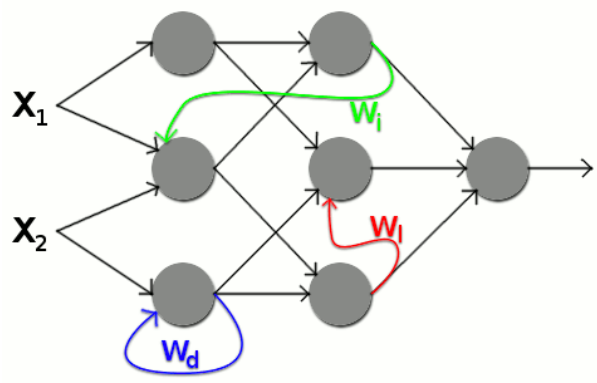

Abbildung 4: Rekurrentes neuronales Netz

Um möglichst viele Szenarien abzubilden, wird das Netz mit unterschiedlichsten Geräuschen trainiert. Hierzu zählen verschiedene Arten von Rascheln, aber auch Geräusche, die dem Rascheln ähneln wie zum Beispiel Karten mischen. Zudem werden Sprachdateien übergeben, die mit verschiedenen Mikrofonen oder in unterschiedlichen Umgebungen aufgenommen wurden. Nach Abschluss des Trainings ist das neuronale Netz in der Lage, Sprache mit Rascheln in reine Sprache und reines Rascheln zu trennen und zu übergeben.[13]

1.4. Fazit und Ausblick

Durch die künstliche Intelligenz haben sich in der Audiopostproduktion einige neue und positive Bearbeitungsmöglichkeiten ergeben. Es wäre ohne künstliche Intelligenz nicht möglich gewesen, Rascheln in diesem Umfang aus Audiodateien zu entfernen, da Rascheln kein kontinuierliches Signal ist. Durch die rasante Entwicklung der künstlichen Intelligenz wird es auch in Zukunft möglich sein, weitere Artefakte zu minimieren und zu entfernen, was aktuell noch nicht möglich ist. Die Software Izotope

[11] Vgl. Izotope, Inc. (2017).
[12] Vgl. Litzel (2017).
[13] Vgl. Izotope, Inc. (2017).

RX 7 Advanced ist neben der Entfernung von Rascheln in der Lage, aus fertig produzierten Musikstücken einzelne Elemente wie die Stimme oder das Schlagzeug herauszufiltern. Diese Funktion ist eine Weiterentwicklung der bereits angesprochenen Quellentrennung. Neben dieser kommt auch Machine Learning zum Einsatz.[14]

Durch den schnellen Fortschritt, der sich durch künstliche Intelligenz ergibt, wird es in einigen Jahren möglich sein, diese Bearbeitungsschritte vollkommen automatisch durchführen zu lassen. Fraglich ist, ob diese Entwicklung nur positive Aspekte erzielt. Möglicherweise sind durch die Automatisierung dieses Vorganges Arbeitsplätze gefährdet. Nichtsdestotrotz sind diese Errungenschaften, die durch künstliche Intelligenz ermöglicht werden, mehr als faszinierend.

[14] Vgl. Robertson & Nercessian (2018).

2. KI und Spracherkennung

In den letzten Jahren hat sich die Spracherkennung rasant entwickelt und bietet inzwischen in nahezu allen Lebenslagen eine Erleichterung der Bedienung von Geräten. Nicht geringer ist das Ziel großer Technologieunternehmen wie Google, Amazon und Co., Maus und Tastatur als Benutzerinterface durch die Spracherkennung zu ersetzen.

Doch was sind die Herausforderungen für eine KI die fehlerfreie Sprache erkennt:

Zum einen muss die Sprache vor allem verstanden werden, dazu gehört auch das korrekte interpretieren des Gesprochenen. Dazu muss die KI die korrekten Suchbegriffe zu den gesprochenen Wörtern liefern. Beispielsweise wenn der Befehl lautet “Spiele Musik von P!NK”, so muss das Ausrufezeichen als “I” interpretiert werden.

Eine weitere Herausforderung besteht darin die Sprache trotz Störgeräusche richtig zu erkennen. So muss das Nutzsignal aus jeglichen Umgebungen extrahiert werden können.

Die letzte Schwierigkeit besteht darin verschiedene Sprachen und Akzente zu implementieren. Jede Sprache funktioniert und klingt anders, genauso gibt es innerhalb von Sprachen, Akzente die sich stark unterscheiden.

Der Vorgang der Spracherkennung lässt sich grob in drei Abschnitte aufteilen, die in den folgenden Kapiteln erläutert werden.

2.1. Analysieren

Folgende Schritte werden am Beispiel der “Amazon Alexa” Erkennung erläutert, unter Umständen unterscheiden sich die Erkennungsverfahren anderer Hersteller.

Das Eingangssignal wird über die Mikrofone aufgenommen und mit 16kHz [15]abgetastet. Da Sprache in einem Bereich von 200 Hz - 5 kHz erkennen lässt, sind nach dem Shannonschen Abtasttheorem[16] alle notwendigen Frequenzen für die Spracherkennung bei einer Abtastung von 16 kHz abgedeckt.

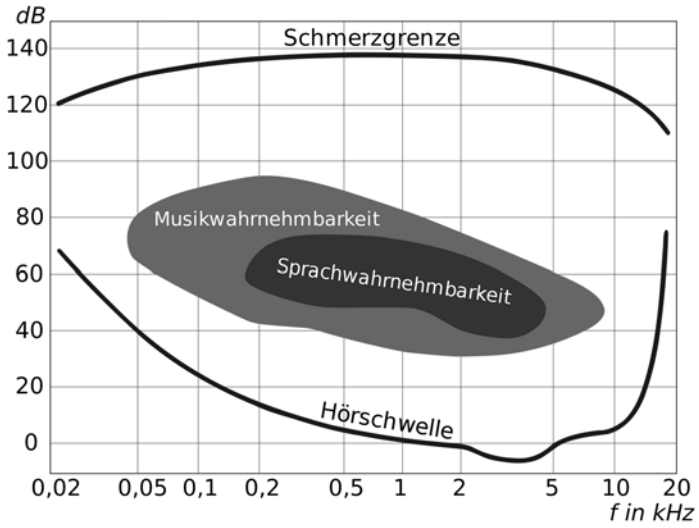

Abbildung 5: Hörfläche des Menschen

Im Anschluss wird das Signal über eine Fourier Transformation in den Spektralbereich umgewandelt und in Spektrale Features unterteilt. Spektrale Features sind eine Unterteilung des Nutzsignals in kleine Bereiche, um die Verarbeitung des Signals im späteren Verlauf zu erleichtern und präzisieren. Diese Features werden dann mithilfe eines Sprachdecoders erkannt und interpretiert. Der komplette Vorgang dauert nur ca. eine Sekunde.

[15] Vgl. Amazon (o.D.).
[16] Vgl. Schwarz (o.D.).

2.2. Verarbeiten

Bei der Aufteilung wird das Signal in sogenannte Phoneme unterteilt. Phoneme sind die kleinste bedeutungsunterscheidende akustische Einheit des Lautsystems einer Sprache. [17] Folgend ein selbst erstelltes Beispiel Signal:

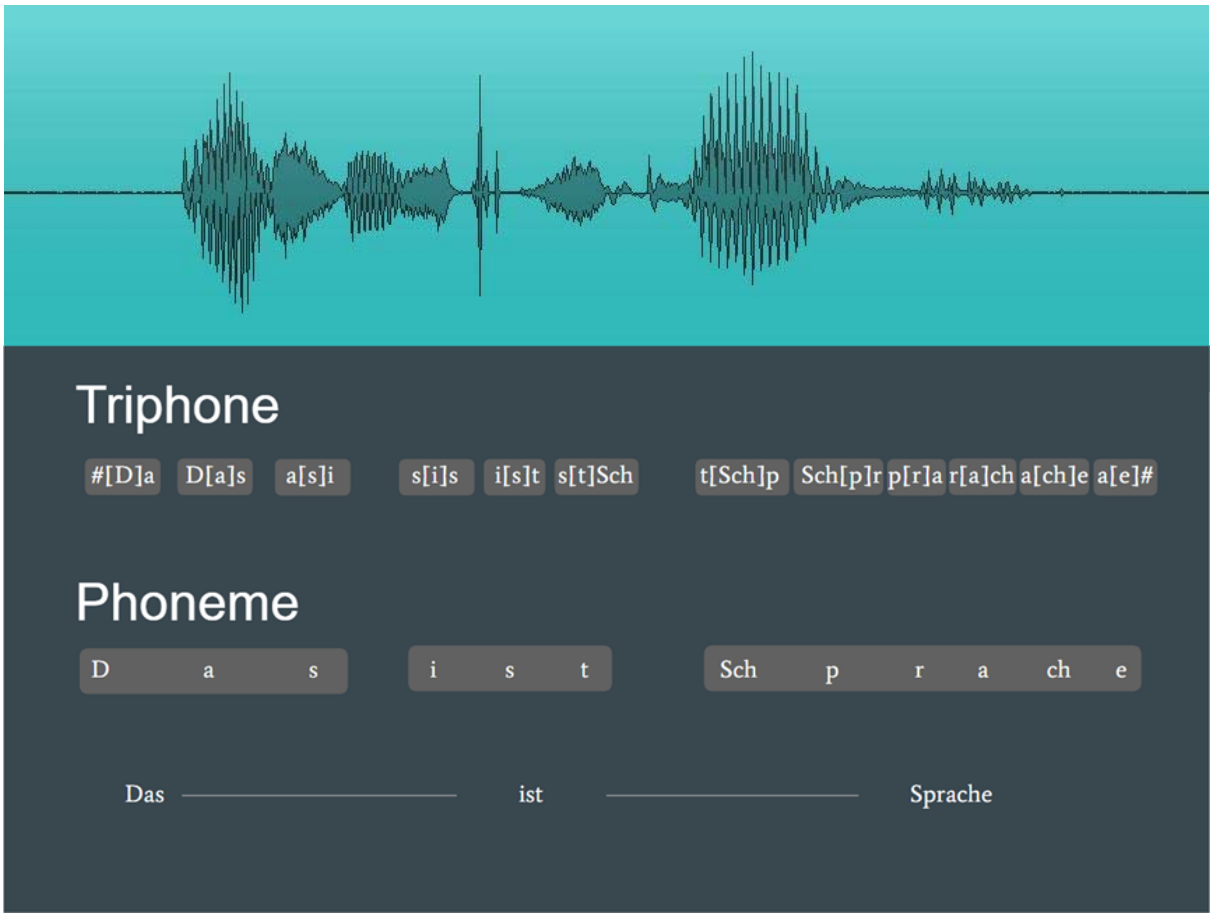

Abbildung 6: Beispielsignal mit zugehörigen Phonemen bzw. Triphonen

Triphone dienen dazu Phoneme in Abhängigkeit ihrer benachbarten Phoneme zu setzen, um auf diese Weise bestimmte Phonem-Kombinationen auszuschließen bzw. mehr in Betracht zu ziehen. Dies funktioniert, da der Klang von Phonemen unter anderem durch vorangegangene und kommende Phoneme beeinflusst wird.[18]

Dieses sog. Akustikmodell unterscheidet sich von Sprache zu Sprache.

[17] Vgl. Duden (2018).
[18] Vgl. CMUSphinx (o.D.).

2.3. Verstehen

Um das Signal korrekt zu interpretieren wird neben dem vorangegangenen Akustikmodell auch ein Sprachmodell angewendet. Aus diesen beiden Teilen und einem Wörterbuch setzt sich der Sprachdecoder zusammen.
In aktuellen Spracherkennungssystemen werden sogenannte “long short term memory” (LSTM) rekursive neuronale Netze (RNN) verwendet.
Diese eignen sich besonders gut zur Modellierung von Sprache, da Sprache eine zeitliche Sequenz ist die zeitlich voneinander abhängige “Werte” hat. Wörter am Satzende können sich z.B. noch auf Wörter vom Satzbeginn beziehen, oder ganze Textabschnitte auf andere.
Außerdem kann je nach Input entschieden werden an wieviel sich das System und dessen Speicher “erinnern” soll. Der Speicher arbeitet dynamisch und löscht selbstständig bereits verstandene und interpretierte Teile des analysierten Sprachsignals, um mehr Speicher für weiteren Input zur Verfügung zu haben.
Hierbei liegt in den USA die Fehlerquote nur nicht bei gut 5%, damit gleicht sie dem Menschen.[19]

2.4. Fazit

Es ist beeindruckend das künstliche Intelligenz in der Spracherkennung den Touring Test bereits bestanden hat. Technisch gesehen ist es lediglich eine Frage der Zeit bis die Spracherkennung auf allen Sprachen fehlerfrei funktioniert. Gleichzeitig sollte man sich bewusst sein, dass in jedem Mobiltelefon und Home-Assistenten ein Mikrofon mit entsprechenden Algorithmen verbaut ist. Theoretisch und technisch sind Unternehmen in der Lage dazu jederzeit die Spracherkennung zu aktivieren und so Gespräche mitzuhören. Selbstverständlich streiten Unternehmen dies ab, sie wären aber in der Lage dazu es zu tun. Auch wenn man die Spracherkennung selbst triggert nimmt das Mikrofon Umgebungsgeräusche auf und erkennt Verhaltensmuster des Nutzers. Es ist wünschenswert, dass von Firmenseite bald eine transparentere und sichere Kultur für die Erhebung von Daten dieser Art geführt wird.

[19] Vgl. ConfEngine (2018).

3. Sprachsynthese durch KI

3.1. Was bedeutet Sprachsynthese?

Neben der im letzten Kapitel näher erläuterten Spracherkennung ist ein weiterer essentieller Aspekt hinsichtlich der menschlichen Interaktion und Kommunikation mit künstlicher Intelligenz die Sprachsynthese. Diese verkörpert die artifizielle Reproduktion einer humanen Sprechstimme. Mittels sogenannter Text-to-Speech-Systeme, kurz TTS, wird dabei ein Fließtext in ein akustisches Signal beziehungsweise in eine akustische Wellenform umgewandelt. Es wird also geschriebene Sprache in Lautsprache gewandelt.[20]

3.2. Etablierte Formen der Sprachsynthese

Konkatenative Sprachsynthese

Bei der konkatenativen Sprachsynthese werden aufgenommene Samples verwendet um Sprache zu erzeugen. Die Sample-Datenbank wird dabei in ganze Sätze, Intonationsphrasen, phonologische Wörter, Silben, Diphone und auch Phoneme unterteilt. Diese unterschiedlichen natürlichen Sprachsegmente werden aneinandergekettet und so ganze Sprachsequenzen gebildet.[21]
Ein großer Nachteil dieses Verfahrens ist, dass das System immer auf die jeweilige Sample-Datenbank beziehungsweise meist auch nur auf einen Sprecher limitiert ist und so nur beschränkt einsetzbar ist.[22] Ein Beispiel für die konkatenative Sprachsynthese ist der Apple-Sprachassistent Siri.

Parametrische Sprachsynthese

Bei der parametrischen Sprachsynthese handelt es sich um vollständig synthetisch hergestellte Sprache. Alle benötigten Daten zur Generierung der gewünschten Sprachsequenzen sind dabei in den Parametern des Modells gespeichert. Die Eigenschaften und der Inhalt der Sprache können über die Eingaben in das Modell gesteuert werden. Mit Hilfe von Signalverarbeitungsalgorithmen, auch bekannt als

[20] Vgl. Pfister & Kaufmann (2008), S. 194.
[21] Vgl. Hande (2014), S. 13.
[22] Vgl. Oloko-oba, Ibiyemi & Samuel (2016), S. 459.

Vocoder, werden bei der parametrischen Sprachsynthese auf diese Weise Sprachsequenzen generiert.[23] Auf diese Weise sind also keine großen Sample-Datenbanken erforderlich.

Der große Nachteil an der parametrischen Sprachsynthese ist, dass die Resultate nicht sehr menschlich, sondern eher künstlich klingen. Ein Beispiel für dieses Verfahren stellt Stephen Hawking's DECtalk Sprachcomputer dar.

3.3. Neuer Ansatz: WaveNet von Google Deepmind

Neben den bereits erläuterten etablierten Formen der Sprachsynthese werden in den letzten Jahren zunehmend neue Konzepte, häufig im Zusammenhang mit künstlicher Intelligenz, entwickelt. Einen interessanten Ansatz verkörpert in diesem Kontext WaveNet von Google Deepmind. Dabei handelt es sich um ein Text-to-speech-System, das mithilfe von KI Audio-Wellenformen synthetisch erzeugen kann. Des weiteren ist das TTS-System so in der Lage jegliche menschliche Stimmen und darüber hinaus sogar sämtliche anderen akustischen Signale relativ authentisch zu imitieren. Wird WaveNet also weiter perfektioniert und verbessert, so ist es in Zukunft möglich akustische Instrumente wie Violinen oder Trompeten, Tiergeräusche, wie ein Löwenbrüllen oder das Bellen eines Hundes, oder jegliche andere Geräusche, wie beispielsweise das eines Automotors, rein synthetisch nachzubilden.
Die Rohwellenform wird von WaveNet direkt modelliert. Dies wird durch ein konvolutionelles neuronales Netzwerk umgesetzt welches aus vielen verschiedenen Faltungsschichten mit diversen Erweiterungsfaktoren besteht und so in der Lage ist mit der Tiefe exponentiell zu wachsen. Auf diese Art und Weise ist es möglich die unzähligen Zeitschritte beziehungsweise Samples eines qualitativ hochwertigen Audiosignals abzudecken.[24] WaveNet ermöglicht also eine schnelle sample basierte Modellierung der Rohwellenform auch bei großen Sampleraten und dementsprechend vielen zu berechnenden Zeitschritten.
In das konvolutionelle neuronale Netzwerk werden in der Trainingsphase echte Aufnahmen menschlicher Sprecher eingespeist. WaveNet erlernt so neben der korrekten Aussprache auch die Stimmeigenschaften des Trainingsmaterials und ist im

[23] Vgl. van den Oord (2016).
[24] Vgl. van den Oord (2016).

Grunde sogar in der Lage Emotionen in der Stimme nachzubilden. Auch Atem- und Schmatzgeräusche sowie Sprachfehler kann das System mit imitieren. Dabei wird für jeden Zeitschritt anhand des vorherigen Samples berechnet welcher Wert auf Basis der Trainings-Wellenformen am wahrscheinlichsten ist und so Schritt für Schritt das Audiosignal berechnet. Dies ist zwar extrem rechenintensiv, erscheint jedoch als notwendig um eine authentisch und realistisch klingende Rohwellenform zu modellieren.[25]

Um qualitativ überzeugende Ergebnisse zu erhalten, muss man WaveNet auch Informationen zu dem Text des zu sprechenden Inhalts geben. Die Vorhersagen des TTS-Systems sind also in der Lernphase nicht nur auf Audio-Samples, sondern auch maßgeblich auf Informationen zu dem Text der eingegeben Wellenformen angewiesen. Der Text wird dabei in eine Reihe phonetischer und sprachlicher Merkmale, wie etwa Informationen zu Phonemen, Silben oder ganzen Wörtern, umgewandelt. Lässt man diese Informationen weg, ist das Netzwerk zwar nach wie vor in der Lage Rohwellenformen zu bilden, die nach einem Menschen klingen, allerdings resultiert dies lediglich in einer undefinierten Abfolge wort ähnlicher Laute.[26]

Eine weitere sehr interessante Eigenschaft des Netzwerks ist außerdem die Tatsache, dass WaveNet, selbst wenn man nur einen einzigen Menschen imitieren möchte, in der Lernphase von Samples verschiedener Sprecher profitiert. Es findet also in gewisser Weise ein Transferlernen statt. Je mehr Material verschiedenster Sprecher also in der Lernphase in das System gespeist wird, desto besser wird das Netzwerk und desto größer wird die Bandbreite der Stimm- und Sprecheigenschaften.[27]

Bei einer Studie der Entwickler von Google Deepmind wurde mittels des MOS-Verfahrens mit der konkatenativen und der parametrischen Sprachsynthese, sowie mit echten menschlichen Sprechern verglichen. MOS steht für Mean Opinion Scores und stellt ein Verfahren zur subjektiven Klangqualitätsprüfung dar, bei der Testpersonen mittels Blindversuchen auf einer Skala von 1 bis 5 eine Bewertung durchführen. Im Hinblick auf die englische und die chinesische Sprache zeigte sich dabei eine deutliche Verbesserung im Vergleich zu den bisherigen Verfahren der Sprachsynthese. Auf Augenhöhe mit einem echten menschlichen Sprecher befindet sich WaveNet jedoch noch nicht.

[25] Vgl. van den Oord (2016)
[26] Vgl. van den Oord (2016)
[27] Vgl. van den Oord (2016)

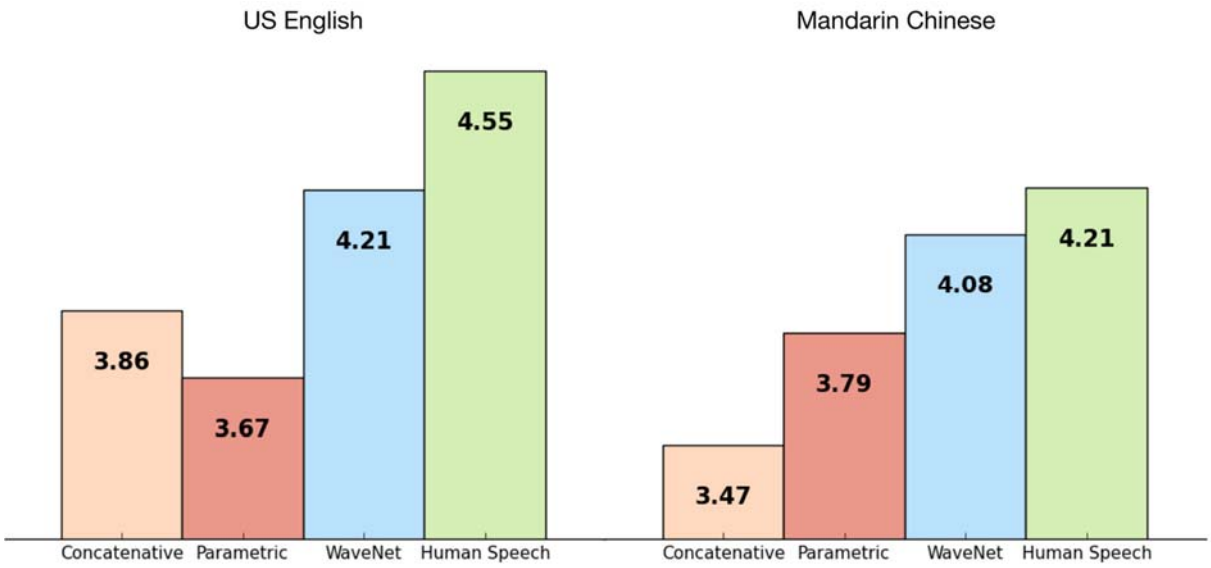

Abbildung 7: Die Ergebnisse des MOS-Verfahrens für die englische bzw. chinesische Sprache graphisch dargestellt

Anhand der Ergebnisse wird also deutlich, dass durch WaveNet und im Allgemeinen durch Sprachsynthese in Kombination mit künstlicher Intelligenz die Lücke zur perfekten Imitation eines menschlichen Sprechers weiter geschlossen wird.

3.4. Fazit

Die technologische Entwicklung der Sprachsynthese wurde durch den Einsatz von künstlicher Intelligenz weiter vorangetrieben. Die bisher etablierten Verfahren, wie die konkatenative oder die parametrische Sprachsynthese, werden mit zunehmender Verbreitung und Massentauglichkeit der neuen Ansätze in Verbindung mit KI langsam verdrängt werden. Dies liegt daran, dass KIs die Modellierung von Rohwellenformen, trotz der enormen erforderlichen Rechenleistung, ermöglichen. Diese klingen qualitativ deutlich besser und ermöglichen des Weiteren die Imitation verschiedener Stimmen, Akzente, Sprachfehler und sogar Atmer bzw. Schmatzer.

Stellt man Überlegungen hinsichtlich der zukünftigen Einsatzgebiete dieser neuen Form der Sprachsynthese an, ergeben sich zunächst offensichtliche Aspekte wie Sprachassistenten, der Einsatz in der Robotik oder auch der Einsatz im Audiobereich. Bei letzterem könnten KIs in Zukunft die Rollen von menschlichen Sprechern und so das Schaffen von Inhalten extrem erleichtern beziehungsweise beschleunigen. Auch ALS-Patienten könnte durch dieses neue Verfahren der Sprachsynthese

bald eine individuelle, authentische Stimme oder gar deren eigene Stimme durch einen Sprachcomputer wiedergegeben werden.
In letzter Zeit machte Sprachsynthese mittels KI allerdings eher durch die Imitation von prominenten Stimmen, wie der des US-Präsidenten Donald Trump auf sich aufmerksam. Hier liegt ein großes Gefahrenpotenzial dieser neuen Technologie: In Zeiten von Fake-News könnte Sprachsynthese durch künstliche Intelligenz auch missbraucht werden, um einflussreichen Menschen falsche Worte in den Mund zu legen und damit eine große Masse zu täuschen. Natürlich wird es bald Software geben, die in der Lage ist manipulierte Stimmen zu erkennen, dennoch gibt es meist kleine Zeitfenster in denen neue Technologien in falschen Händen großen Schaden anrichten können.
Die Sprachsynthese im Zusammenhang mit künstlicher Intelligenz ist also durchaus auch kritisch zu sehen. Dennoch birgt sie große Chancen und viel Potenzial, die Menschheit und besonders ihre Interaktion mit Computern technologisch voran zu bringen.

4. Abbildungsvereichnis

5. Quellenverzeichnis

Amazon. (o.D.). *SpeechRecognizer Interface.* Abgerufen am 25.02.2019 unter https://developer.amazon.com/de/docs/alexa-voice-service/speechrecognizer.html#recognize

Beck, F. (2018). *Neuronale Netze – Eine Einführung – Units.* Abgerufen am 22.12.2018 unter http://www.neuronalesnetz.de/units.html

CMUSphinx. (o.D.). *Basic concepts of speech recognition.* Abgerufen am 25.02.2019 unter https://cmusphinx.github.io/wiki/tutorialconcepts/

ConfEngine. (06.09.2018). *The Deep Learning Revolution in Automatic Speech Recognition by Dr Ananth Sankar at #ODSC_India.* Abgerufen am 25.02.2019 unter https://www.youtube.com/watch?v=oAhe4DmbygM

Duden. (o.D.). *Pho-nem, Fo-nem*. Abgerufen am 25.02.2019 unter https://www.duden.de/rechtschreibung/Phonem

Hande, S. (2014). *A Review of Concatenative text To Speech Synthesis*. EXTC Department, Navi Mumbai.

Haselhuhn, A. (12.01.2018). Rekurrente / rückgekoppelte neuronale Netzwerke. Abgerufen am 12.12.2018 unter https://dbs.uni-leipzig.de/file/Haselhuhn_Folien.pdf

iZotope, Inc. (24.04.2017). *DeRustle: Removing Lavalier Microphone Noise with Deep Learing*. Abgerufen am 24.11.2018 unter https://techblog.izotope.com/2017/04/24/derustle-removing-lavalier-microphone-noise-with-deep-learning/

Kruse, R., Borgelt, C., Braune, C., Klawonn, F., Moewes, C., & Steinbrecher, M. (2015). *Computational Intelligence: Eine methodische Einführung in Künstliche Neuronale Netze, Evolutionäre Algorithmen, Fuzzy-Systeme und Bayes-Netze* (2. Aufl.). Wiedbaden, Deutschland: Springer Fachmedien Wiesbaden.

Litzel, N. (26. April 2017). *Was ist Deep Learning?.* Abgerufen am 22.12.2018 unter https://www.bigdata-insider.de/was-ist-deep-learning-a-603129/

Moeser, J. (27. September 2017). *Künstliche Neoronale Netze – Aufbau & Funktionsweise.* Abgerufen am 22.12.2018 unter https://jaai.de/kuenstliche-neuronale-netze-aufbau-funktion-291/

Oloko-oba, M.; Ibiyemi T.S, I.; Samuel, O. (2016). *Text-to-Speech Synthesis Using Concatenative Approach. International Journal of Trend in Research and Development*. Volume 3. S. 459-462.

Pfister, B.; Kaufmann, T. (2008). *Sprachverarbeitung - Grundlagen und Methoden der Sprachsynthese und Spracherkennung.* Zürich, Schweiz: Springer-Verlag Berlin Heidelberg.

Robertson, H., & Nercessian, S. (25. September 2018). *Exploring the Technology that Makes RX 7 Music Rebalance Possible.* Abgerufen am 07.01.2019 unter https://www.izotope.com/en/blog/music-production/exploring-the-technology-that-makes-rx-7-music-rebalance-possible.html

Schwarz, A. (o.D.). *Abtasttheorem.* Abgerufen am 25.02.2019 unter https://www.mikrocontroller.net/articles/Abtasttheorem

van den Oord, A. (8. September 2016). *Wavenet: A Generative Model for Raw Audio.* Abgerufen am 23.02.2019 unter https://deepmind.com/blog/wavenet-generative-model-raw-audio/