TECHNISCHE UNIVERSITÄT
BERGAKADEMIE FREIBERG

Die Ressourcenuniversität. Seit 1765.

Fakultät für Mathematik und Informatik
Institut für Informatik
Professur für Künstliche Intelligenz und Datenbanken

Seminararbeit

Automatische Spracherkennung am Beispiel Deep Speech

Max Mustermann

Masterseminar Angewandte Informatik - Trends in Deep Learning
Matrikel: 11111

1. April 2021

Inhaltsverzeichnis

Motivation

Der Einsatz von Systemen zur automatischen Spracherkennung hat in den letzten Jahren stark zugenommen. Dabei dienen entsprechende Systeme oft der Assistenz bei der Erfüllung von Aufgaben und dem Komfort. Automatische Spracherkennung ist die Grundlage für eine Sprachsteuerung, die es ermöglicht Kommandos und Eingaben freihändig aufzugeben. Der Einsatzbereich solche Erkennungs- und Steuersysteme reicht von häuslichen Anwendungen bis hin zur hochmodernen Industrie. Auch im Bereich der digitalen Barrierefreiheit gewinnen Spracherkennungssystem und Sprachsteuerungen an immer größerer Bedeutung für beeinträchtigte Nutzer. Die Anforderungen an die Systeme sind aber in allen verschiedenen Anwendungsbereichen gleich. Die Systeme sollen so genau wie möglich arbeiten und robust gegenüber der Umgebung sein, in der sie zum Einsatz kommen. Zudem ist eine leichte Adaption eines bestehenden Systems in andere Sprachen sehr wünschenswert, um sie für alle Nutzer zugänglich zu machen.
Die Seminararbeit entstand im Rahmen des Masterseminars "Trends in Deep Learning" im Sommersemester 2020 an der TU Bergakademie Freiberg und gibt einen Überblick über die aktuelle Umsetzbarkeit der beschriebenen Anforderungen an automatische Spracherkennungen. Im ersten Kapitel werden grundlegende Probleme bei automatischer Spracherkennung beschrieben. Im zweiten Kapitel wird auf das quelloffene Spracherkennungssystem "Deep Speech" eingegangen und die enthaltenen Komponenten werden erläutert. Zudem wird ein Vergleich zur Leistungsfähigkeit anderer Spracherkennungssysteme gezogen. Im dritten Kapitel werden Möglichkeiten beschrieben, das betrachtete System auf neue Sprachen zu trainieren.

1 Grundlagen

Um das Problem der automatischen Spracherkennung zu lösen, sind zwei verschiedene Modelle bisher erfolgreich angewandt wurden. Bereits in den 1970er Jahren wurden erstmals Hidden-Markov-Modelle eingesetzt, um menschliche Sprache in Maschinentext zu überführen. Hidden-Markov-Modelle simulieren einen zweistufigen stochastischen Prozess, der Ergebnisse über Wahrscheinlichkeitsfunktionen errechnet. Für die Spracherkennung ist dabei die Wahrscheinlichkeit ausschlaggebend, mit der ein Laut auf einen anderen folgen kann.

Der zweite Ansatz nutzt künstliche neuronale Netze. Auch wenn die Theorie dieser Netze ebenfalls schon Mitte des 20. Jahrhunderts entwickelt wurde, erlebten die Netze eine Renaissance mit der Verfügbarkeit hoher Rechenleistungen Anfang der 2000er Jahre. Die dadurch resultierende Möglichkeit extrem große Datenmengen in einer überschaubaren Zeit verarbeiten zu können, befeuerte das Forschungsgebiet erneut. Diese Seminararbeit beschäftigt sich mit dem Einsatz solcher künstlichen neuronalen Netze.

1.1 Probleme der automatischen Spracherkennung

Bei der Auseinandersetzung mit Systemen zur automatischen Spracherkennung gibt es grundlegende Problemstellungen, von welchen alle Systeme gleichermaßen betroffen sind und die es stets zu berücksichtigen gilt.

Schlechtes Signal-Rausch-Verhältnis

Das Signal-Rausch-Verhältnis beschreibt die Qualität des zu bearbeitenden Signals. Im Falle von automatischer Spracherkennung ist dies die Aufnahme der gesprochenen Sprache, welche in Text umgewandelt werden soll. Das Nutzsignal, ist hierbei die menschliche

Sprache. Als Störsignal fließen alle Hintergrundgeräusche und mögliches Rauschen ein, die die eigentliche Aufnahme zum Teil überlagern. Darunter fallen neben Umgebungsgeräuschen wie Musik oder Verkehrslärm auch technische Ursachen, wie schlechte Aufnahmetechnik. Je höher der Pegel des Nutzsignals gegenüber des Störsignals ist, umso besser lässt die Aufnahme sich verarbeiten.

Akzente/Dialekte
In jeder Sprache gibt es regionsabhängige Variationen der eigentlichen Amtssprache. Besonders Betonungen von Wörtern und Verwendung verschiedener Grammatiken erschweren die Arbeit automatischer Spracherkennungssysteme.

Natürliche Sprache
Bei einer natürlichen Konversation halten sich Konversationsteilnehmer nicht immer an die grammatikalischen Regeln einer Sprache. Sätze werden beispielsweise nicht ganz vollendet oder es wird eine Menge von Füllwörtern benutzt.

Lombard-Effekt
Dieser Effekt beschreibt das Phänomen, dass Menschen ihre Stimmfrequenz automatisch anpassen, um auftretende Störgeräusche in ihrer Umgebung zu kompensieren. Der Lombard-Effekt lässt sich nicht aktiv steuern und muss bei automatischer Spracherkennung mit beachtet werden.

Sprechgeschwindigkeit
Jeder Mensch hat seine eigene Sprechgeschwindigkeit, welche auch von Situation zu Situation variabel ist. Unterschiedlich lang gesprochene Phrasen können dasselbe ausdrücken. Demgegenüber müssen Spracherkennungssysteme robust sein.

Die beschriebenen Probleme treten alle im Praxiseinsatz eines automatischen Spracherkennungssystems auf. In einer kontrollierten Aufnahmeumgebung, bei der sich die Sprecher auf ihre Aufgabe korrekt zu sprechen konzentrieren können, treten diese Probleme nur sehr selten auf. Da die Systeme aber alle zum Ziel haben in natürlichen Umgebungen und als Assistenzsystem zu funktionieren, müssen die Aspekte vor allem im Trainingsprozess mit beachtet werden.

1.2 Word Error Rate

Eine Möglichkeit automatische Spracherkennungssysteme nach ihrer Leistungsfähigkeit und Genauigkeit bei der Erkennung menschlicher Sprache zu bewerten stellt die "Word Error Rate" (WER) dar. Diese Fehlergröße ist in der Literatur die am häufigsten genutzte und wird auch beim Einsatz von Benchmarks oft angewendet. Diese Fehlerrate beschreibt den prozentualen Anteil der Fehler im Ergebnis der Spracherkennung im Bezug zum eigentlich gesprochenen Satz. Dabei werden immer ganze Worte betrachtet.
Die Word Error Rate ergibt sich aus der Anzahl der gesprochenen Worte und den aufgetretenen Fehlern. Als Substitution wird dabei eine Ausgabe bezeichnet, welche zwar korrekt als zusammenhängendes Wort erkannt wurde, jedoch nicht dem tatsächlich gesprochenen Wort entspricht (beispielsweise wird "Baum" fälschlicherweise als "Traum" ausgegeben).

$$WER = \frac{S + D + I}{N} \tag{1}$$

mit:

S = Anzahl der Substitutionen
D = Anzahl fehlender Worte
I = Anzahl eingefügter Worte
N = Anzahl der gesamten Worte

Eine niedrige Fehlerrate beschreibt in dem Fall ein sehr genaues System. Diese Art der Bewertung von Sprachsystemen wird teilweise kritisiert, da bei diesem Bewertungssystem kein Unterschied gemacht wird inwieweit ein Wort falsch verstanden wurde oder gar nicht. Auch wenn der eigentliche Sinn der gesprochenen Sprache in der durch die Spracherkennung erstellten Transkription noch erhalten ist, fließt eine Ungenauigkeit als voller Fehler in die Bewertung durch die Word Error Rate ein. Dies tritt beispielsweise auf, wenn grammatikalische Endungen falsch interpretiert werden. Wird anstatt "ich gehe" die Transkription "ich gehe*n*" ausgegeben, gilt dies als voller Fehler, obwohl das Wort korrekt erkannt wurde und nur ein Buchstabe in den Endung fehlerhaft ist.
Eine andere gängige Möglichkeit der Bewertung von automatischen Spracherkennungssystemen ist die "Phenom Error Rate" (PER). Hier wird bereits während der Erkennung einzelner Laute oder Buchstaben die Fehlerrate gemessen. Bei vielen Spracherkennungssystemen entspricht dies aber nicht dem tatsächlichen Ergebnis, welches am Ende ausgegeben wird, da meist noch ein Abgleich über ein Sprachmodell erfolgt (vgl. Kapitel 2.3).

1.3 Trainingsprozess und Trainingsdaten

Um ein künstliches neuronales Netz auf eine bestimmte Anwendung zu trainieren, sind eine große Anzahl aufbereiteter Trainingsdaten notwendig. Das sogenannte "Deep Learning" basiert darauf, dass einem künstlichen neuronalen Netz während eines überwachten Trainingsprozesses Beispieldaten zugeführt werden, für die das zu erwartende Ergebnis bereits bekannt ist. Für das Trainieren von automatischer Spracherkennung benötigt man transkribierte Audio-Daten. Dies sind Sprachaufnahmen, für die das gesprochene Wort bereits in elektronischer Schriftform vorliegt. Um solche Daten zu generieren sind zwei Methoden gängig.

1. Aufzeichnung von Konversationen und nachträgliches Transkribieren durch Menschen
2. Aufzeichnung von vorgelesenen Texten

Einfacher und schneller gestaltet sich das Vorlesen von Texten, da hier die Transkription bereits als Ausgangspunkt vorliegt. Allerdings werden bei dieser Methode keine realen Gesprächsbedingungen abgebildet, wie es bei Konversationen der Fall ist. Bei Letzteren sorgt das händische Transkribieren aber für zusätzlichen Aufwand.
Es gibt für viele Sprachen bereits frei zugängliche Datenbanken mit Trainingsdaten. Für weniger oft gesprochene Sprachen sind auch die Datensätze bedeutend kleiner, was einen Trainingsprozess für diese Sprachen sehr erschwert.

1.4 Digitalisierung der Sprache

Um das gesprochene Wort zu digitalisieren wird im ersten Schritt die Schallwechseldruckschwingung, welche beim Sprechen entsteht und das gesprochene Wort für Menschen hörbar macht, mittels eines Mikrofons in ein elektrisches Signal umgewandelt. Durch den Einsatz eines Analog-Digital-Wandlers wird dieses Signal dann in diskrete Zustände übersetzt, die von Computern verarbeitbar sind. Hierbei wird jeweils zu einem Zeitpunkt die entsprechende Amplitude des Signals abgespeichert. Für die Anwendung der automatischen Spracherkennung muss dieses Signal aber noch weiter aufbereitet werden, da die Information über die

Stärke eines Signals zu einem bestimmten Zeitpunkt keine Informationen über das was zu diesem Zeitpunkt gesprochen wurde bereitstellt. Um ein verwertbares Signal zu erhalten, muss die Frequenzverteilung innerhalb eines betrachteten Zeitabschnittes des Signals bekannt sein. Eine solche Frequenzanalyse lässt sich durch den Einsatz einer Fast-Fourier-Transformation erhalten. Dadurch wird der Anteil der verschiedenen Frequenzen innerhalb eines Zeitabschnitts errechnet, was für eine Analyse der Sprache notwendig ist. Ein so aufbereitetes Signal kann dann im nächsten Schritt an ein künstliches neuronales Netz weitergegeben werden.

2 Deep Speech

Als quelloffenes Projekt wurde "Deep Speech" im Jahr 2014 von der Mozilla Foundation der Öffentlichkeit erstmalig vorgestellt [1]. Als Alternative zu kommerziellen Produkten großer Anbieter war das System bereits 2014 in der Lage in Benchmarktests mit diesen zu konkurrieren. Besonderes Augenmerk lag bei der Entwicklung des Systems auf Spracherkennungen in schwierigen Situationen (spontane Konversation, viele Hintergrundgeräusche, etc.). Nachfolgend werden die wichtigsten Hauptkomponenten des Systems vorgestellt.

2.1 Rekurrentes neuronales Netz (RNN)

Um die aufbereiteten Sprachaufnahmen zu analysieren und eine Abschätzung der zum jeweils betrachteten Zeitpunkt geäußerten Laute vornehmen zu können, setzt Deep Speech auf eine relativ überschaubare Architektur seines verwendeten neuronalen Netzes. Schematisch ist das RNN in Abbildung 1 abgebildet.

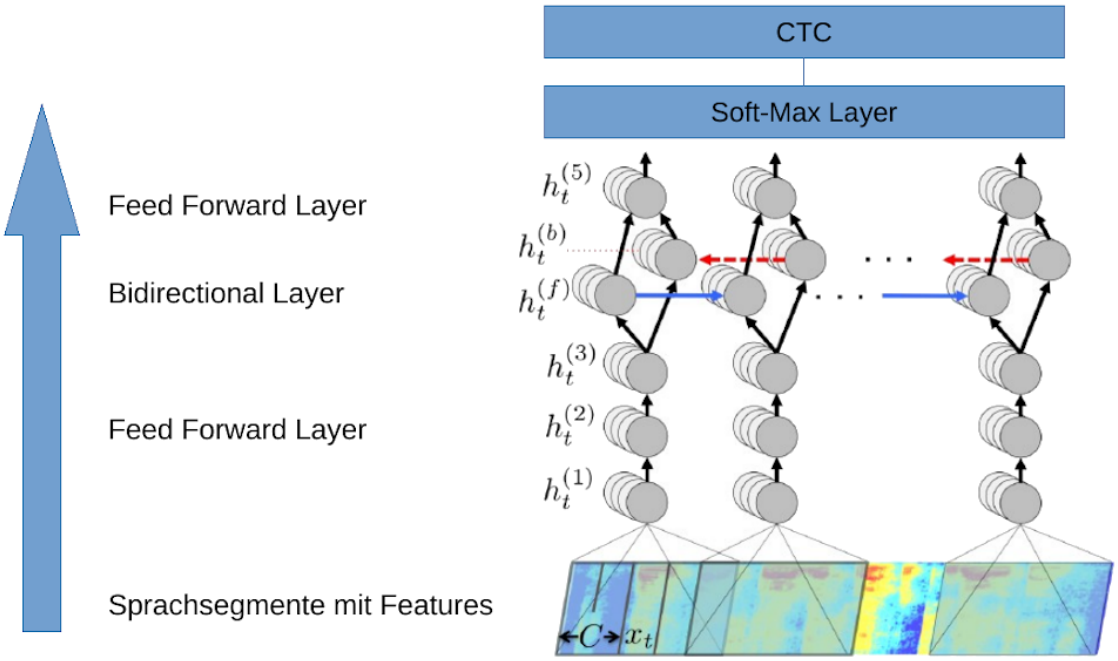

Abb. 1: Schematische Darstellung des RNN mit Anschluss an das nachfolgende Element CTC [1]

Als Eingabe für den Input-Layer werden die einzelnen Frequenzspektrogramme der zu analysierenden Audio-Datei verwendet. Dafür werden immer Zeitbereiche von 20 Millisekunden als ein Zeitabschnitt betrachtet [2]. In Abbildung 1 sind diese Zeitabschnitte mit x_t gekennzeichnet. Neben der Möglichkeit nur einen einzelnen Zeitabschnitt als Input zu verwenden wurde auch versucht zusätzlich zu dem jeweiligen betrachteten Abschnitt x_t die Vorgänger und Nachfolger Spektren als Kontext mit in das Netzwerk zu füttern. Je nachdem wie groß dieser Kontextbereich gewählt ist, wird dieser mit C bezeichnet ($C \in [0, 3, 5, 9]$). In der Praxis stellte sich der am größten gewählte Bereich $C = 9$ als am günstigsten heraus.

Die Hidden-Layer des Netzes besteht aus nur fünf verschiedenen Ebenen $h_t^{(1\text{-}5)}$. Davon sind vier als Feedforward Layer konzipiert und nur der Layer an vierter Stelle ist ein Bidirectional Layer. Nur in dieser Ebene kommt es zum Austausch zwischen den einzelnen Neuronen derselben Ebene. Eine zu analysierende Audio-Datei wird hier komplett betrachtet und die Informationen von jedem Zeitabschnitt der Datei hier über das komplette Netzwerk abgeglichen. Die geschieht einmal in Vorwärtsrichtung $h_t^{(f)}$ und einmal in Rückwärtsrichtung $h_t^{(b)}$. Nachdem dieser Abgleich vollzogen wurde, wird das Ergebnis über den Outputlayer ausgegeben. In der ersten Version von Deep Speech (2014) wurde absichtlich auf den Einsatz von Long-Short-Term-Memory Zellen (LSTM) verzichtet, da man hohe Einbußen in der benötigten Rechenzeit befürchtete. Ab Version 2 (2015) wurden im Bidirektionalen Layer LSTMs eingesetzt. [2]

Als Output-Layer wird ein Soft-Max-Layer verwendet. Für das entsprechende Alphabet der verwendeten Sprache wird die Wahrscheinlichkeit für die Äußerung jedes möglichen Buchstabens c ($c \in [a, b, c, ..., z, Leerzeichen]$) für jeden betrachteten Zeitabschnitt x_t als $\vec{y}_t = P(c|x_t)$ ausgegeben.

2.2 Connectionist Temporal Classification (CTC)

Durch die natürliche Sprache der Menschen sind gesprochene Äußerungen immer von unterschiedlicher Länge, auch wenn das Gleiche gesagt wird. Ein und dieselbe Äußerung kann von unterschiedlichen Menschen oder in unterschiedlichen Situationen gesprochen eine sehr unterschiedliche Länge durch andere Sprechgeschwindigkeiten oder Betonungen aufweisen. Dementsprechend kann der Output des Soft-Max-Layer für ein und dasselbe Wort sehr unterschiedliche ausfallen:

	Zeitabschnitt :	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}	x_{t6}	x_{t7}	x_{t8}	x_{t9}	x_{t10}
Fall 1	$y_t = P_{max}(c\|x_t)$:	h	a	l	_	o	o	_	_	_	_
Fall 2	$y_t = P_{max}(c\|x_t)$:	h	a	a	l	l	o	o	o	o	_

In beiden Fällen ist dasselbe Wort gesprochen worden, allerdings in unterschiedlichen Geschwindigkeiten. Diese unterschiedlichen Eingaben müssen aber auf dieselbe Ausgabe gebracht werden. Um dieses Problem zu lösen wird der "Connectionist Temporal Classification"-Algorithmus (CTC) verwendet. Hier werden betrachtete Zeitabschnitte x_t, die zu der selben Äußerung eines Lautes beziehungsweise eines gesprochenen Buchstabens gehören zusammengefasst. Damit wird es möglich die Ausgabe des neuronalen Netzes von aufeinanderfolgenden Lauten auf zusammenhängende Wörter abzubilden, wodurch eine korrekte Spracherkennung und ein Trainingsprozess des Netzwerkes möglich wird. Die Ausgabe des CTCs lässt sich mit den erwarteten Ergebnissen der Trainingsdaten abgleichen und so ein korrekt ermittelter Buchstabe von einem falschen Ergebnis unterscheiden. Erst durch dieses Wissen ist es möglich mittels Backpropagation einen Lernprozess innerhalb des künstlichen neuronalen Netzes durchzuführen.

2.3 Sprachmodell

Die Ergebnisse, welche durch das RNN und CTC geliefert werden, weisen bereits sehr gute Ergebnisse auf, betrachtet man die erkannten Buchstaben jeweils für sich. Bei dem Vergleich von ganzen Wörtern mit den Originaldaten kommt es aber noch häufig zu Fehlern, da einzelne Buchstaben falsch erkannt sind und damit das ganze Wort nicht korrekt ist. Die Fehlerursache

hierfür liegt zumeist in der Tatsache begründet, dass die Aussprache bestimmter Worte nicht mit ihrer Schreibweise übereinstimmt und das System eher dazu tendiert eine "Lautschrift" des gesprochenen Wortes auszugeben und nicht die korrekte Rechtschreibung. Um diesem Problem entgegenzuwirken, wird bei Deep Speech ein Sprachmodell eingesetzt. Für die englische Sprache steht ein 5-gram Sprachmodell mit einem Vokabular von 495.000 Wörtern zur Verfügung. Mit über 220 Millionen englischen Phrasen wurde das Modell so trainiert, dass es in der Lage ist semantische Zusammenhänge und Wahrscheinlichkeiten für Wortgruppen - bestehend aus dem bekannten Vokabular - von einer Länge bis zu 5 Wörtern zu erfassen.
Die Wörter, welche von RNN/CTC ausgegeben werden, werden anschließend in das Sprachmodell geben. Hier wird über eine Wahrscheinlichkeitsfunktion abgewogen, ob das erkannte Wort dem Vokabular entspricht oder korrigiert werden sollte. Mathematisch wird das Maximum der Funktion über die angegeben Wahrscheinlichkeiten für einen Buchstaben aus dem neuronalen Netz und dem Sprachmodell gesucht.

$$Q(c)=log(P(c|x)) + \alpha * log(P_{\mathrm{lm}}(c)) + \beta * word\_count(c)_{[1]} \tag{2}$$

Alpha und Beta sind dabei Gewichtungsfaktoren, die die Wahrscheinlichkeit aus dem Sprachmodell P_{lm} höher wichten, je länger die untersuchte Wortgruppe ist.

2.4 Trainingsdaten

Für das Training von Deep Speech wurde zum einen auf frei zugängliche Trainingsdaten in englischer Sprache zurück gegriffen. Der Trainingsdatensatz "Switchboard" liefert circa 300 Stunden aufgenommene und transkribierte Konversation. Der Datensatz "Fisher"besteht sogar aus über 2000 Stunden transkribiertem Audiomaterial. Zum anderen wurde von dem mit "Baidu" gemeinsam betriebenen Projekt "Common Voice" noch einmal 5000 Stunden Audiomaterial von freiwilligen Nutzern gesammelt. Diese Daten sind allerdings keine Konversationen, sondern vorgelesene Texte. Insgesamt standen somit 7000 Stunden Trainingsdaten in englischer Sprache zur Verfügung.

Um die Menge der Trainingsdaten künstlich zu erweitern wurden diese absichtlich mit Störgeräuschen versehen. Dafür wurden kurze Audiospuren mit Störgeräuschen zufällig über die bereits vorhanden Audioaufnahmen gelegt. Ebenfalls wurde der Lombard-Effekt bei einigen Aufnahmen künstlich erzeugt, indem die Sprecher bei der Aufnahme der Audiodaten über Kopfhörer zufällig Störgeräuschen ausgesetzt wurden.

2.5 Vergleich zu anderen Spracherkennungssystemen

Damit ein Vergleich über die Genauigkeit von Deep Speech zu anderen automatischen Spracherkennungen möglich ist, wurde ein Benchmarktest herangezogen, sowie ein durch die Entwickler von Deep Speech eigens entwickelter Testdatensatz eingesetzt.

2.5.1 Hub500 Benchmark

Der eingesetzte "Hub500 Benchmark" besteht aus 40 aufgezeichneten Telefongesprächen. Bei der Hälfte dieser Gespräche war den Konversationspartner ein Thema und ein Skript vorgegeben. Dieser Datensatz wird als "Switchboard" (SWB) bezeichnet. Die andere Hälfte der Konversationen sind spontane Gespräche ohne vorgegebenes Thema. Dieser Datensatz heißt "Call Home" (CH). Auf Grund der Spontanität der Dialoge in diesem Datensatz gilt

dieser als der anspruchsvollere für automatische Spracherkennungssysteme.

Um einen verlässlichen Vergleich zu ziehen wurde für diesen Test Deep Speech nur mit den 2300 Stunden Trainingsdaten trainiert, welche die frei zugänglichen Datensätze "Switchboard" und "Fisher" zur Verfügung stellen. Das als State-of-the-Art geltende automatische Spracherkennungssystem, mit dem 2014 verglichen wurde, stammt von Mass et al. [3] und wurde nur mit dem "Fisher"-Datensatz (2000 Stunden) trainiert. Die Ergebnisse sind in Tabelle 1 aufgeführt. Mittlerweile werden für den Hub500 Benchmark auch noch bessere Ergebnisse erzielt. Eine Übersicht der aktuell führenden Systeme ist in Tabelle 2 gegeben.

Modell	SWB	CH	Gesamt
Maas et al. (FSH)	16.0	23.7	19.9
Deep Speech (SWB + FSH)	12.6	19.3	16.0

Tab. 1: Ergebnis in % WER. Benchmark von 2014 [1][3]

Modell	SWB	CH	Gesamt
CAPIO (2017) [4]	5.0	9.1	7.1
Highway LSTM (2017) [5]	5.1	9.9	7.5
Microsoft (2016) [6]	6.3	11.9	9.1
Deep Speech (2014) [1]	12.6	19.3	16.0

Tab. 2: Ergebnis in % WER. Aktuelle Benchmarks

2.5.2 Vergleich verrauschter Aufnahmen

Die Entwicklung von Deep Speech hat zum Ziel auch Sprache, welche in sehr lauten Umgebungen aufgenommen wurde, möglichst fehlerfrei zu erfassen. Um dies zu überprüfen und einen Vergleich mit bereits existierenden Spracherkennungssystemen schaffen zu können, wurde ein eigener Benchmarktest entwickelt. Dieser besteht aus je 100 rauschfreien und 100 verrauschten Aufnahmen, die von insgesamt 10 unterschiedlichen Sprechern eingesprochen wurden. Für diesen Test wurde Deep Speech mit den über 7000 Stunden vorhandenem Trainingsmaterial trainiert, welches bis zu zwanzigmal mit immer neuen Störgeräuschen versetzt wurde und dann wieder als neuer Trainingsdatensatz zur Verwendung kam. Damit kam man auf eine Gesamtlänge von über 100.000 Stunden Audiomaterial, was dem Spracherkennungssystem vor dem Test zugeführt wurde.
Für das Benchmark wurden noch drei weitere Systeme mit denselben Daten getestet und anschließend auch für diese die Word Error Rate bestimmt. Konnte ein Spracherkennungssystem für eine bestimmte Aufnahme gar kein Ergebnis liefern, so wurde die Aufnahme komplett aus dem Test entfernt, um eine Vergleichbarkeit trotzdem weiterhin zu gewährleisten. Von den 200 Sprachaufnahmen, die für das Benchmark aufgenommen wurden, flossen dadurch nur 176 tatsächlich ein. Eine Übersicht über die Ergebnisse des Tests ist in Tabelle 3 gegeben.

Durch die Ergebnisse wird für mich gut ersichtlich, dass besonders bei den verrauschten Aufnahmen das System Deep Speech bedeutend bessere Ergebnisse erzielt. Damit sehe ich das

System	kein Rauschen	verrauscht	kombiniert
Apple Dictation	14.24	43.76	26.73
Bing Speech (2017)	11.73	36.12	22.05
Google API	6.64	30.47	16.72
wit.ai	7.94	35.06	19.41
Deep Speech	6.56	19.06	11.85

Tab. 3: Ergebnis in % WER. Eigenes Benchmark Deep Speech [1]

nachträgliche Anreichern von Aufnahmen mit synthetisierten Störgeräuschen als eine praktikable Möglichkeit an, um Systeme robuster gegenüber akustischen Störeinflüssen zu machen.

3 Aktuelle Problemstellungen

3.1 Neue Sprachen

Für ein robustes Spracherkennungssystem ist es wichtig, dass es schnell und einfach auf andere Sprachen anpassbar ist. Hierfür werden aber in jedem Fall Trainingsdatensätze in der jeweiligen gewünschten Sprache, sowie ein entsprechendes Sprachmodell benötigt. Für Sprachen, die anteilsmäßig wenig gesprochen werden, sind meist auch nur wenige Trainingsdaten verfügbar. Trotz dessen soll Deep Speech in der Lage sein, auch mit wenigen Trainingsdaten für neue Sprachen schnell akzeptable Ergebnisse liefern zu können. Im Nachfolgenden werden drei Versuche dahingehend zusammen mit den Ergebnissen präsentiert.

3.1.1 Mandarin

Den ersten Versuch Deep Speech für eine andere Sprache als Englisch zu verwenden, unternahmen die Entwickler von Deep Speech selbst. Die Ergebnisse dazu wurden im Paper "Deep Speech 2: End-to-End Speech Recognition in English und Mandarin"[2] im Dezember 2015 veröffentlicht. Für Mandarin stand eine große Menge an Trainingsdaten zur Verfügung. Insgesamt wurden für die Versuche Sprachaufnahmen im Umfang von 9.400 Stunden verwendet, die über 11.000 verschiedene Äußerungen enthielten. Als Sprachmodell kam ein umfangreiches 5-gram Modell zum Einsatz, welches über 2 Milliarden 5-grams verfügte.

Um die Ergebnisse des Systems in Mandarin zu messen, wurde zum Einen ein Test mit 2.000 Äußerungen aus den vorhandenen Trainingsdaten verwendet. Hier erzielte das System eine WER von nur 5.81 %. Weiterhin wurde noch ein Test mit 1.882 verrauschten Aufnahmen durchgeführt, die nicht im Trainingsprozess mit eingesetzt wurden. Hier erreichte man immer noch eine WER von 7.93 %. Im Zusammenhang mit diesen Tests wurde auch die Fehlerquote von Menschen, die einen Text transkribieren sollen, ermittelt. Für 100 zufällig gewählte Äußerungen transkribierte eine Testgruppe von fünf Personen mit einer Fehlerrate von 4.0%. Für dieselben Äußerungen konnte das trainierte System sogar mit nur 3.7% WER aufwarten.

3.1.2 Russisch

2018 unternahm man an der St. Petersburg Universität den Versuch das quelloffene System Deep Speech auf die russische Sprache anzupassen. Hier standen bedeutend weniger Trainingsdatensätze als in vorherigen Versuchen zur Verfügung. Circa 1000 Stunden Audiomaterial und zugehörige Transkription wurden russischen YouTube-Videos entnommen. Die Transkriptionen waren teilweise direkt von Nutzern erstellt und teilweise automatisch generiert, was schon ein gewisses Fehlerpotential in den Trainingsdaten birgt. Ergänzt wurden die Daten noch mit einem Datensatz von 650 Stunden sauber eingesprochen und transkribiertem Material. Auch wurde ein Sprachmodell mit implementiert, welches mit den Daten russischer Wikipedia-Seiten erstellt wurde. Welche Dimension das Sprachmodell aufwies, wurde allerdings nicht angegeben.

Hier machte es keinen großen Unterschied, ob das System mit den transkribierten YouTube-Videos (1000 Stunden - WER 27%) oder den sauberen Aufnahmen (650 Stunden - WER 28%) trainiert wurde. Das beste Ergebnis ließ sich erzielen, indem man beide Datensätze als Trainingsdaten für das System verwendete. Hiermit erreichte man eine WER von 22% auf dem verwendeten Testsatz, welcher allerdings nur saubere Aufnahmen enthielt. [7]
Aus den Ergebnissen schlussfolgere ich, dass die Qualität der Spracherkennung signifikant von der Menge der verfügbaren Trainingsdaten abhängig ist. Der erreichte Wert ist 2018 für saubere Aufnahmen schon nicht mehr zeitgemäß (vgl. Tabelle 2). Weiterhin wurde in der Veröffentlichung selbst noch einmal die Bedeutung des Sprachmodells hervorgehoben. Ohne den Einsatz eines Sprachmodells fielen die Ergebnisse mit 10% mehr absoluter WER aus [7].

3.2 Unüberwachtes Lernen

Die vorhergehenden Betrachtungen haben gezeigt, dass die Qualität einer automatischen Spracherkennungssoftware stark von der Menge verfügbarer Trainingsdaten abhängt. Gerade beim Erlernen neuer Sprachen stellt dies oft ein Problem dar, da für viele Sprachen keine großen Datensätze an transkribierten Audioaufnahmen verfügbar sind. Eine Lösung für dieses Problem wurde 2019 von Yi Ren et al. im Paper "Almost Unsupervised Text to Speech and Automatich Speech Recognition" vorgestellt [8].
Der Kerngedanken, der in dieser Arbeit verfolgt wird, stellt die Dualität zwischen den Aufgaben automatische Spracherkennung und automatische Sprachgenerierung dar. Während beim Text-to-Speech (TTS) aus vorhandenem Text Sprache generiert wird, wandelt eine automatische Spracherkennung (ASE) Sprache wieder in Text um. Stellt man einen vorgegebenen Text als x dar, so kann die zugehörige Sprachausgabe aus dem TTS als $\hat{y}$ geschrieben werden. Ein Datenpaar lässt sich damit als $(x, \hat{y})$ darstellen. Für die ASE kann $\hat{y}$ dann als Eingabe dienen, die den ursprünglichen Text wieder abbilden soll $(\hat{y}, \hat{x})$. Über die Abweichung von x und $\hat{x}$ lässt sich ein Fehler ermitteln, der dann für den Lernprozess der beiden Systeme benutzt werden kann.

Für die Tests mit dem entworfenen System wurde der frei zugängliche Datensatz "LJSpeech" verwendet, welcher rund 24 Stunden transkribiertes Audiomaterial in englischer Sprache bereitstellt. Um die Effizienz des entworfenen Systems abschätzen zu können wurden drei verschiedene Tests durchgeführt. Die beiden Systeme wurden einmal mit dem kompletten Material trainiert (Supervised Trainig). Im zweiten Durchlauf wurden den jeweiligen Systemen nur 200 transkripierte Aufnahmen (etwa 20 Minuten) zum Training gegeben (Pair-200). Während

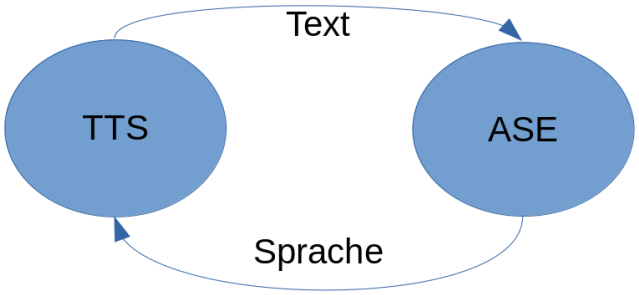

Abb. 2: Dualität TTS und ASE

beim Supervised Training TTS wie auch ASE beide sehr gute Ergebnisse erzielen konnten, konnte bei einem Training mit Pair-200 keine erkennbare Sprache durch das TTS erzeugt werden und die Fehlerrate der ASE fiel auch extrem hoch aus. (vgl. Tabelle 4)
Im dritten Versuch wurde wieder der Pair-200 Datensatz als Trainingsdaten für die beiden Systeme verwendet. Daran schloss sich dann ein weiterer Trainingsprozess an, der das entworfene System und die Dualität der beiden Aufgaben ausnutzte. Die übrigen Datensätze von LJSpeech, die nicht im Pair-200 enthalten waren, wurden aufgeteilt in nur Sprache und nur Text und dann den jeweiligen Systemen als Eingabe gegeben, mit dem Ziel sich gegenseitig zu trainieren (Almost Unsupervised Training). Auch wenn die Ergebnisse dieser Methode nicht an die des Supervised Trainings herankamen, waren sie doch bedeutend besser, als die Pair-200 Methode, welche nur mit 20 Minuten Trainingsmaterial auskommen musste.

Methode	**PER**
Supervised	2.5%
Pair-200	72.5%
Almost Unsupervised	11.7%

Tab. 4: Ergebnisse der drei verschiedenen Ansätze in %PER [8]

Damit lässt sich für Sprachen, die über keine großen Mengen aufbereiteter Trainingsdaten verfügen, eine Möglichkeit schaffen auch mittels nicht transkribierten Eingabedaten ein funktionierendes System zu schaffen.

4 Fazit

Systeme zur automatischen Spracherkennung sind bereits jetzt in der Lage Spracheingaben mit sehr hohen Genauigkeiten zu transkribieren. Auch in schwierigen Umgebungen mit einer hohen Anzahl von Störeinflüssen können mittlerweile beachtliche Ergebnisse erzielt werden. Als Problem bleibt weiterhin die große Menge an benötigten Trainingsdaten bestehen, die für jede Sprache neu zur Verfügung gestellt werden muss. Hier zeigt sich für mich vor allem in den Ergebnissen der Arbeit aus St. Petersburg (vgl. Kapitel 3.1.2), dass die Qualität der automatischen Spracherkennung sehr von der Menge der Daten abhängt und weniger von deren Qualität.
Um diesem Problem Abhilfe zu schaffen, finde ich die Ansätze aus den Arbeiten zu Deep Speech [1] [2] interessant, bereits bestehende Aufnahmen künstlich mit Störeinflüssen zu versetzen, um so Systeme robuster gestalten zu können und automatisch mehr Trainingsdaten zu generieren. Die Ergebnisse aus Kapitel 2.5.2 zeigen auf, dass dieser Ansatz auch funktioniert. Einen Schritt weiter geht das in Kapitel 3.2 beschriebene System, bei dem die Dualität von TTS und ASE ausgenutzt wird. Ich sehe in diesem Ansatz sehr großes Potenzial, um vor allem Spracherkennungen mit akzeptablen Ergebnissen für wenig gesprochene Sprachen oder Dialekte trainieren zu können. Texte in digitaler Form liegen in großen Mengen für fast alle Sprachen zugänglich vor. Ebenso sind Audiomitschnitte oder Videos mit Tonspuren eher verfügbar als transkribiertes Audio. Mit dem beschriebenen Ansatz ließe sich die Menge der benötigten Daten in aufbereiteter Form so weit reduzieren, dass Systeme in kurzer Zeit auf neue Sprachen trainiert werden könnten.
Überraschend war für mich die Tatsache, wie hoch der Einfluss des verwendeten Sprachmodells auf die Ergebnisse der automatischen Spracherkennung mittels neuronalen Netzen ausfällt. Sprachmodelle beschreiben die Wahrscheinlichkeit von zueinander gehörenden oder aufeinander folgenden Wörtern. Damit sind sie eng verwandt mit den Hidden-Markov-Modellen. Neuronale Netze scheinen für mich die Hidden-Markov-Modelle zur automatischen Spracherkennung dementsprechend nicht zu ersetzen, sondern vielmehr zu ergänzen.

Literatur

[1] Hannun A., et al, *Deep Speech: Scaling up end-to-end speech recognition*, arXiv:1412.5567, 2014.

[2] Hannun A., et al, *Deep Speech2: End-to-End Speech Recognition in English and Mandarin*, arXiv:1512.02595, 2015.

[3] Maas A., et al, *Increasing deep neural network acoustic model size for large vocabulary continuous speech recognition*, arXiv:1406.7806, 2015.

[4] Kyu H. Han., et al, *The CAPIO 2017 Conversational Speech Recognition System*, arXiv: 1801.00059v2, 2018.

[5] Kurata G., et Al, *Language Modeling with Highway LSTM*, arXiv: 1709.06436, 2017.

[6] Stolcke A., et Al, *The Microsoft 2017 Conversational Speech Recognition System*, arXiv:1708.06073v2, 2017.

[7] Degtyarev A., et Al, *Russian-Language Speech Recognition System Based on Deepspeech*, 2014.

[8] Ren Y., et Al, *Almost Unsupervised Text to Speech and Automatic Speech Recognition*, arXiv:1905.06791, 2019.