Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Hidden-Markov-Modelle als statistisches Mittel zur Spracherkennung: Ein Vergleich zwischen Laut- und Wort-Modellen

Verfasser	:	Thore Nitz
Matrikelnummer	:	7301409
Studiengang	:	Software-System-Entwicklung
E-Mail	:	thore.nitz@studium.uni-hamburg.de

Dozent	:	Dr. Timo Baumann
Abgabedatum	:	11. September 2020

Abstract

Um mit der Variabilität von Sprachsignalen bestmöglich umgehen zu können, hat sich die stochastische Modellierung zur Spracherkennung mit Hilfe von Hidden-Markov-Modellen (HMM) als geeignete Methode herausgestellt. Um mit Hilfe von HMM einen Spracherkenner zu modellieren, gibt es zwei unterschiedliche Ansätze. Zum einen den Ansatz der Wort-HMM, bei dem jeweils pro Modell ein einzelnes Wort beschrieben wird, und zum anderen den der Laut-HMM, bei dem im Gegensatz dazu deutlich kleinere linguistische Einheiten wie einzelne Phoneme oder Triphone beschrieben werden. Bei Betrachtung der Komplexität und Flexibilität dieser beiden Ansätze hat sich über die Jahrzehnte der Ansatz der Laut-HMM als deutlich effektiverer und praxisrelevanterer Ansatz durchgesetzt.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

Soll eine gesprochene Sprache mit Hilfe eines elektronischen Geräts erkannt und anschließend in einen Text umgewandelt werden, so ist zunächst das Vorhandensein von Sprachsignalen vonnöten. Diese entstehen, wenn eine Person, die durch die Sprache produzierten Schallwellen bspw. in ein Mikrofon spricht, welches dann als Folge die akustischen in elektrische Signale umwandelt. Die daraus entstehenden Sprachsignale werden beeinflusst durch viele verschiedene Faktoren und sind damit weitestgehend unterschiedlich. So verändern die Emotionen, die Aussprache oder gar die Physiologie der Sprecherin oder des Sprechers die gesprochenen Laute und damit auch die daraus entstehenden elektrischen Signale. [1, S. 25 f.] Um mit dieser Variabilität von Sprachsignalen bestmöglich umgehen zu können, hat sich neben weiteren Möglichkeiten der Spracherkennung die stochastische Modellierung mit Hilfe von HMM als eine geeignete Methode herausgestellt.

HMM sind in der Spracherkennung eine weitverbreitete Technik und beruhen auf dem Prinzip der Markov-Ketten, einem stochastischen Prozess, bei dem es um die Wahrscheinlichkeit des Eintretens zukünftiger Ereignisse geht. [1, S. 109] Um mit Hilfe von HMM eine Spracherkennung zu modellieren, gibt es zwei unterschiedliche Ansätze. Die Wort-HMM beschreiben jeweils pro Modell ein einzelnes Wort und geben bei einer Eingabesequenz sodann das am wahrscheinlichsten passende Modell in Form eines Wortes als Ausgabe. [1, S. 341] Um kontinuierliche Sprache besser zu erkennen, gibt es ebenfalls den Ansatz der Laut-HMM. Hierbei werden nicht einzelne Worte in Form eines HMM repräsentiert, sondern deutlich kleinere linguistische Einheiten wie einzelne Phoneme. [1, S. 363]

In der folgenden Seminararbeit wird in *Kapitel 2* zunächst das HMM als stochastisches Modell vorgestellt und seine Struktur, die einzelnen Parameter sowie die drei grundlegenden Probleme näher beschrieben. Anschließend befasst sich das *Kapitel 3* mit der Verwendung der HMM zur statistischen Spracherkennung. Hierbei werden sowohl die Wort-HMM als auch die Laut-HMM vorgestellt und die Funktionsweise dieser näher beschrieben. Nachdem die Vor- und Nachteile beider Ansätze in *Kapitel 4* diskutiert wurden, fasst das *Kapitel 5* die Seminararbeit mit einem Fazit und einer Handlungsempfehlung zusammen.

2 Hidden-Markov-Modelle

Bei HMM handelt es sich grundlegend um stochastische Modelle, um bspw. in der Spracherkennung verschiedene Sprachsignale analysieren und verarbeiten zu können. Die Basis der HMM sind Markov-Ketten, welche im folgenden *Kapitel 2.1* zunächst als Grundlage näher beschrieben werden, damit anschließend in *Kapitel 2.2* sowie *Kapitel 2.3* näher auf die Struktur und Funktionsweise der eigentlichen HMM eingegangen werden kann.

2.1 Markov-Ketten

Eine Markov-Kette ist ein Modell, welches Aussagen über die Wahrscheinlichkeit von Sequenzen unterschiedlicher Zustände einer Menge möglich macht. Diese Mengen können dabei verschiedene Dinge repräsentieren wie das Wetter, Symbole oder im Falle der Spracherkennung einzelne Laute oder gar gesamte Wörter. Bei der Verwendung von Markov-Ketten ist insbesondere hervorzuheben, dass bei einer Vorhersage über die Zukunft einer sequenziellen Zustandsfolge lediglich der aktuelle Zustand ausschlaggebend ist und einer historischen Betrachtung keinerlei Einfluss für die Zukunft zugesprochen wird. Würde man nun bspw. eine Vorhersage über das Wetter von morgen treffen wollen, so ist dann das Wetter von heute ausschlaggebend, nicht jedoch das von gestern. [2, S. 149 f.]

Eine Markov-Kette besteht aus einer endlichen Menge an Zuständen Q, den Übergangswahrscheinlichkeiten zwischen den einzelnen Zuständen A sowie der Startwahrscheinlichkeit S. In der folgenden *Abbildung 1* werden die Bestandteile einer Markov-Kette weitergehend definiert sowie anhand eines Beispiels näher erläutert. [2, ebd.]

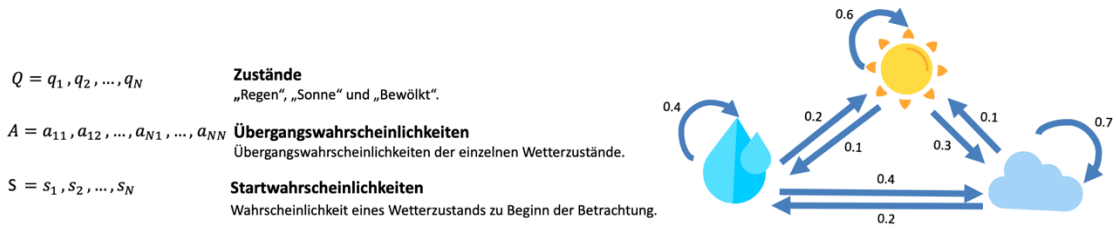

Abbildung 1: Definition und Beispiel einer Markov-Kette. Eigene Darstellung, vgl. [2, ebd.]

Möchte man nun also, ausgehend vom Wetter heute, eine Aussage über die Wahrscheinlichkeit des Wetterverlaufs in der Zukunft tätigen, ist dies anhand des in *Abbildung 1* gezeigten, fiktiven Beispiels möglich.

2.2 Struktur und Parameter der Hidden-Markov-Modelle

Wie im vorherigen *Kapitel 2.1* beschrieben, lässt sich mit Hilfe der Verwendung von Markov-Ketten die Wahrscheinlichkeit von außen sicht- und beobachtbaren Zuständen errechnen. In vielen Fällen sind diese eigentlich beobachtbaren Zustände jedoch verborgen (engl. hidden). Stattdessen sind jedem dieser inneren Zustände beobachtbare Ausgabesymbole, sogenannte Emissionen, zugeordnet, die je nach Zustand mit gewissen Wahrscheinlichkeiten auftreten. Die Aufgabe besteht nun darin, aus der beobachteten Sequenz der Emissionen zu wahrscheinlichkeitstheoretischen Aussagen über die verborgenen Zustände zu kommen. Anhand eines Beispiels lässt sich diese Besonderheit anschaulich erklären: Wollen wir bspw. die Wortarten innerhalb eines gegebenen Textes zuordnen, so können wir dies meist nicht direkt, ohne den gesamten Kontext zu betrachten und mit einzubeziehen. Wir sehen vielmehr die Wörter und müssen diesen die Wortarten aus der Wortfolge heraus ableiten. Die Wortarten sind bei diesem Beispiel zunächst verborgen, weil sie nicht direkt beobachtet werden können und die Wörter die jeweiligen sichtbaren Ausgabesymbole bzw. Emissionen. [2, S. 150 f.]

Um nun ein HMM genauer definieren zu können, müssen zur bisher bekannten Definition von Markov-Ketten aus *Abbildung 1* noch die Ausgabesymbole O sowie die Emissionswahrscheinlichkeiten B hinzugefügt werden. In der folgenden *Abbildung 2* werden die Bestandteile eines HMM definiert sowie erneut anhand eines Beispiels erläutert.

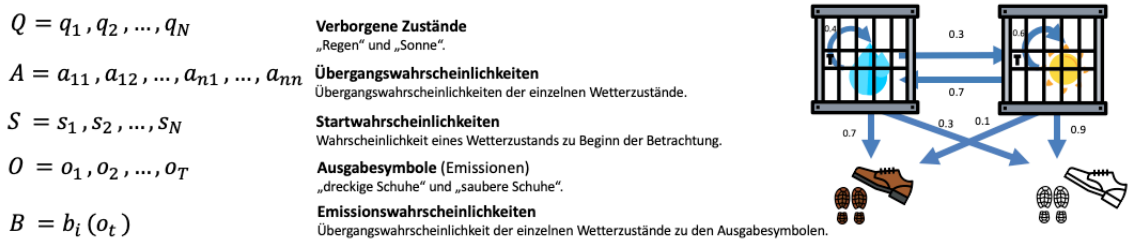

Abbildung 2: Definition und Beispiel eines HMM. Eigene Darstellung, vgl. [2, ebd.]

Betrachtet man nun die Beschaffenheit der Schuhe als sichtbare Ausgabesymbole, also ob diese dreckig oder sauber sind, so lässt sich folglich ein Rückschluss auf die Wahrscheinlichkeit des Wetters, in Form der verborgenen Zustände, formulieren.

Überträgt man diesen Aufbau nun auf die automatische Spracherkennung basierend auf HMM, so werden die gesprochenen Laute oder Wörter als versteckte Zustände aufgefasst und die tatsächlich hörbaren Töne als Emissionen. [1, S. 336 i.V.m. S. 363]

2.3 Die grundlegenden Probleme der Hidden-Markov-Modelle

Bei Betrachtung der im vorherigen *Kapitel 2.2* beschriebenen Form der HMM ergeben sich drei grundlegende Probleme, die es zu lösen gilt. Spricht man in diesem Zusammenhang von Problemen, so beschreibt dies keine negativen Aspekte der HMM, sondern vielmehr Fragestellungen, durch dessen Lösungen das Modell einen realen Mehrwert erzeugt. Die drei Probleme lauten *Evaluationsproblem*, *Decodierungsproblem* sowie *Schätzproblem* und werden im Folgenden näher beschrieben und erläutert. Im Rahmen dieser Seminararbeit wird dabei auf detaillierte Erläuterungen der mathematischen Vorgehensweise verzichtet. [3, S. 8]

Evaluationsproblem

Das Evaluationsproblem behandelt die Fragestellung, wie man anhand eines gegebenen HMM sowie einer gegebenen Sequenz von Ausgabesymbolen die Wahrscheinlichkeit errechnen kann, mit der die beobachtete Sequenz von ebendiesem HMM produziert wurde. Dies bietet zum einen die Möglichkeit, ein einzelnes HMM zu bewerten und zum anderen aus einer Menge von HMM eine Art Ranking aufzustellen, welches dieser die Ausgabesequenz am wahrscheinlichsten produziert hat und welches am unwahrscheinlichsten. Hierbei wird von der sogenannten Produktionswahrscheinlichkeit gesprochen. Eine Lösung für das Evaluationsproblem bietet der sogenannte Forward-Algorithmus. [3, ebd.]

Decodierungsproblem

Das Problem, eine Aussage über die verborgenen Zustände eines HMM ziehen zu können, nennt man Decodierungsproblem. Hierbei soll durch eine gegebene Sequenz von Ausgabesymbolen diejenige Sequenz an verborgenen Zuständen herausgearbeitet werden, die die Beobachtungen am wahrscheinlichsten produziert hat. Dadurch ist es bspw. möglich, die Struktur des Modells kennen zu lernen und Statistiken sowie das Verhalten innerhalb einzelner Zustände näher zu beschreiben. Betrachten wir das Beispiel in *Abbildung 2,* so lässt sich bei Lösung des Decodierungsproblems bspw. erläutern, welches Wetter bei der jeweiligen Beschaffenheit der Schuhe am wahrscheinlichsten ist. Einen effizienten Weg, um das Decodierungsproblem zu lösen, bietet der Viterbi-Algorithmus. [1, S. 114]

Schätzproblem

Das dritte Problem, das Schätzproblem, befasst sich mit der Optimierung eines HMM, sodass die einzelnen Modellparameter so bestimmt werden, dass das HMM eine

Sequenz von Ausgabesymbole möglichst gut beschreibt. Ziel ist es also, anhand von Trainings-Beobachtungssequenzen die Parameter des HMM so zu bestimmen, dass die Produktionswahrscheinlichkeit (siehe Evaluationsproblem) für das jeweilige Beispiel maximiert wird. [1, ebd. f.] Das Training von HMM ist eine entscheidende Voraussetzung, um Modelle für real existierende Phänomene zu erstellen. Eine Lösung für das Schätzproblem bietet der Baum-Welch-Algorithmus. [3, S. 8]

Die HMM werden als statistisches Modell erfolgreich in der Spracherkennung eingesetzt. Wie dies funktioniert und welchen praktischen Nutzen dabei die Modelle bieten, wird im folgenden *Kapitel 3* näher behandelt. Nach einer allgemeinen Einführung wird zunächst das Wort-HMM in *Kapitel 3.1* und darauf aufbauend das Laut-HMM in *Kapitel 3.2* vorgestellt und näher beschrieben.

3 Statistische Spracherkennung

Um in der Spracherkennung die Variabilität von Sprachsignalen bestmöglich verarbeiten zu können, bietet die statistische Spracherkennung mit Hilfe von HMM eine geeignete Lösung. Hierbei kann die Erkennung als solche als Decodierungsproblem (siehe *Kapitel 2.3*) oder Evaluationsproblem interpretiert werden. [1, S. 327 i.V.m S. 342] Die folgende *Abbildung 3* zeigt beispielhaft den Ablauf einer statistischen Spracherkennung aus informationstheoretischer Sicht und der Erkennung mit Hilfe des Decodierungsproblems.

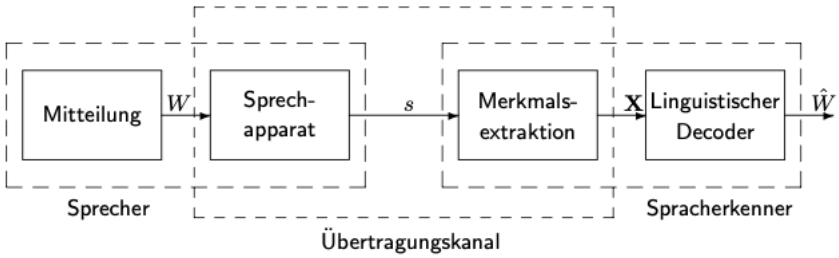

Abbildung 3: Spracherkennung aus informationstheoretischer Sicht. [1, S. 327]

Zunächst werden die gesprochenen Worte W mit Hilfe eines Sprechapparats in die akustischen Sprachsignale s umgewandelt. Nach der Übertragung dieser Signale zum eigentlichen Spracherkenner wird anhand einer Merkmalsextraktion eine Sequenz verschiedener Merkmale X extrahiert. Diese Sequenz stellt im Bezug zu den HMM die beobachtbare Sequenz an Ausgabesymbolen dar. X lässt sich als Ausgabe des Übertragungskanals definieren, in den initial W als Mitteilung gegeben wurde. Der linguistische Decoder hat nun die Aufgabe aus der Merkmalssequenz die eigentlich geäußerte

Wortfolge zu ermitteln, welche sich hier als die verborgenen Zustände eines HMM definieren lassen. Dies bedeutet, mit Hilfe des Decodierungsproblems eine möglichst gute Schätzung der Wortfolge $\hat{W}$ zu ermitteln. Um dies zu tun, wird aus stochastischer Sicht innerhalb eines großen HMM zu der gegebenen Merkmalssequenz X diejenige Wortfolge $\hat{W}$ gesucht, welche von allen Folgen von Wörtern eines gegebenen Vokabulars V mit der kleinsten Wahrscheinlichkeit zu einem Fehlentscheid führt. Dieses vorgehen wird auch *Maximum-a-posteriori-Regel* genannt und wie folgt definiert: $\hat{W} = \underset{W \in V^*}{argmax}\ P(W|X)$. [1, S. 327 f.] Gibt es nun verschiedene HMM, die bspw. jeweils eine andere Wortfolge $\hat{W}$ beschreiben, so wird im Gegensatz zu vorherigem Beispiel mit Hilfe des Evaluationsproblems das HMM ermittelt, welches die höchste Produktionswahrscheinlichkeit der Merkmalssequenz X aufweist. Die Wortfolge $\hat{W}$, die dieses ausgewählte HMM dann repräsentiert, wurde am wahrscheinlichsten vom Sprecher geäußert und gilt damit als erkannt. [1, S. 342]

Um abschließend die drei in *Kapitel 2.3* beschriebenen Probleme der HMM und dessen Verwendung in der statistischen Spracherkennung besser nachvollziehen zu können, lassen sie sich diese erneut anhand eines Schemas näher erklären. Möchte man bspw. jeweils ein HMM mit N verschiedenen verborgenen Zuständen für jedes Wort[1] eines Vokabulars mit V Worten entwerfen, müssen diese zunächst trainiert werden. Man beginnt nun also mit einer Trainingssequenz für jedes Wort V, welche aus einer Anzahl von Wiederholungen des gesprochenen Wortes besteht. Mit Hilfe des Baum-Welch-Algorithmus des Schätzproblems wird als Folge für jedes Wort des Vokabulars ein HMM mit optimalen Modellparametern erstellt. Um jede Trainingssequenz in verschiedene Zustände zu segmentieren, damit ein Verständnis für die innere Struktur der HMM zu bekommen und die HMM noch weiter trainieren zu können, verwenden wir die Lösung des Decodierungsproblems. Um nun im letzten Schritt unbekannte Worte zu erkennen, kann schließlich der Forward-Algorithmus des Evaluationsproblems genutzt werden. Das HMM, welches die größte Produktionswahrscheinlichkeit bei Eingabe eines unbekannten Sprachsignals widergibt, beschreibt das eingegebene Wort. [3, S. 8] Eine Alternative zur Erkennung anhand vieler einzelner HMM wäre, wie oben bereits beschrieben, die Erkennung mit Hilfe einer Decodierung. [1, S. 327]

[1] Anstelle von Worten können hier natürlich auch einzelne Laute verwendet werden. Die genaue Differenzierung findet jedoch erst im folgenden *Kapitel 3.1* und *Kapitel 3.2* statt.

Die Spracherkennung mit Hilfe von HMM unterscheidet zwischen zwei verschiedenen Ansätzen. Zum einen den Wort-HMM und zum anderen den Laut-HMM. Beide Ansätze werden im folgenden *Kapitel 3.1* bzw. *Kapitel 3.2* näher erläutert.

3.1 Wort-Hidden-Markov-Modelle

Werden zur Erstellung eines Spracherkenners für jedes zu erkennende Wort jeweils ein HMM erzeugt, so spricht man von einer Spracherkennung mit Hilfe von Wort-HMM. Grundlegend gibt es bei dieser Art von Erkennung drei Möglichkeiten der Umsetzung, welche im Folgenden jeweils kurz vorgestellt werden.

Bei einem *Einzelworterkenner* wird für jedes Wort ein unabhängiges HMM erzeugt, welches ebendies repräsentiert. Schaut man sich bspw. eine Spracherkennung für Ziffern an, so gibt es jeweils für „null“, „eins“, ..., „neun“ ein eigenes HMM. Die Ermittlung des gesprochenen Wortes kann somit mit Hilfe des Vergleichs der Produktionswahrscheinlichkeiten (siehe Evaluationsproblem) und damit mit dem Forward-Algorithmus erfolgen. Alternativ, und ein in der Praxis relevanterer Ansatz, ist jedoch auch die Erkennung anhand des Viterbi-Algorithmus (siehe Decodierungsproblem). Hierbei werden dann folglich nicht mehr die Produktionswahrscheinlichkeiten miteinander verglichen, sondern die Verbundwahrscheinlichkeiten der Merkmalssequenz und der optimalen Zustandssequenz. [1, S. 342 f.]

Bleibt man bei der Erkennung mit Hilfe des Viterbi-Algorithmus, so ist es möglich, einen *Spracherkenner mit Erkennungsnetzwerk* zu erzeugen. Bei dieser Möglichkeit werden bspw. die Ziffern des oben beschriebenen Beispiels jeweils mit einem gemeinsamen Anfangs- und Endzustand zusammengefasst. Dieser Vorgang wird als Parallelschaltung der vorher eigenständigen Wort-HMM bezeichnet. Durch diesen Aufbau sorgen wir dafür, dass die einzelnen Ziffern zu den verborgenen Zuständen des Erkennungsnetzwerk werden und eine Erkennung mit Hilfe des Decodierungsproblems möglich wird. [1, S. 343]

Eine letzte Möglichkeit des Aufbaus eines Wort-HMM bietet die *Verbundworterkennung*. Durch diese ist es möglich, nicht nur einzelne gesprochene Wörter nacheinander zu erkennen, sondern ganze Folgen dieser. Um einen solchen Vorgang zu ermöglichen, werden zunächst alle zu erkennenden Folgen von Wörtern in Serie geschaltet. Anschließend werden diese Folgen aufbauend auf die Spracherkennung mit Erkennungsnetzwerk mit Hilfe eines gemeinsamen Anfangs- und Endzustandes parallel

miteinander verbunden. Ein Beispiel für die praktische Umsetzung einer Verbundworterkennung wäre bspw. die Erkennung von Vorwahlen. Hierbei wird zunächst jede mögliche Vorwahl anhand ihrer Ziffern in Serie geschaltet und diese anschließend miteinandener verbunden. Die eigentlich Erkennung funktioniert dann erneut mit dem Viterbi-Algortihmus. [1, S. 345 ff.]

3.2 Laut-Hidden-Markov-Modelle

Anstatt für jedes einzelne Wort ein HMM zu erzeugen, können diese auch kleinere linguistische Einheiten repräsentieren. Die kleinsten Grundelemente zur Erkennung mit Hilfe eines HMM wären dann bspw. Phoneme oder Triphone. Werden Phoneme mit Hilfe von HMM repräsentiert, so spricht man im Zuge der Spracherkennung von kontextunabhängigen Grundelementen, da für jedes Phonem nur ein zugehöriges HMM existiert. Berücksichtigt man jedoch, dass ein Phonem von seinen Nachbarlauten beeinflusst wird, werden die Grundelemente kontextabhängig betrachtet und bspw. für jedes Triphon ein HMM erzeugt. [1, S. 349] Schaut man sich diese Vorgehensweise anhand eines Beispiels an, wird deutlich, warum der Kontext einzelner Phoneme mitbetrachtet werden sollte, damit eine umfängliche Spracherkennung möglich gemacht wird. Betrachtet man das Phonem *[n]*, so verändert sich das akustische Signal dieses bei Aussprache in unterschiedlichen Konstellationen wie *[ana]*, *[ini]* oder *[ang]*. [1, S. 352 f.]

Hat man nun eine Zusammenstellung mehrerer trainierter Laut-HMM, lässt sich aus diesen ein Spracherkenner für verschiedenste Einsatzgebiete konstruieren. Sollen mit Hilfe der Laut-HMM zunächst nur einzelne Wörter erkannt werden, ist dieser ähnlich aufgebaut wie der oben bereits beschriebene Einzelworterkenner, nur dass zuvor die einzelnen Laute des Wortes zu einem Verbund-HMM zusammengesetzt werden müssen. Anschließend erfolgt die Vernetzung dieser Verbund-HMM zu einem Erkennungsnetzwerk, sodass einzelne Wörter dann mit Hilfe des Viterbi-Algorithmus erannt werden. [1, S. 363]

Um hingegen kontinuierliche Sprache mit Hilfe von Laut-HMM möglich zu machen, bedarf es zwei weiterer Komponenten. Zum einen ist ein *Aussprache-Lexikon* vonnöten, welches alle zu erkennenden Wörter anhand ihrer phonetischen Umschrift aufgelistet hat und zum anderen ein *Sprachmodell*, welches die Wahrscheinlichkeiten bestimmter Wortfolgen verzeichnet. Aus allen vorhandenen Laut-HMM wird sodann ein *Erkennungsnetzwerk* zusammengestellt, sodass für die eigentliche Erkennung der kontinuierlich gesprochenen Sprache der Viterbi-Algorithmus eingesetzt werden kann. [1, S. 364]

4 Vor- und Nachteile beider Ansätze

Beide in *Kapitel 3* vorgestellten Ansätze zur Erkennung von Sprache mit Hilfe von HMM als statistische Herangehensweise lassen sich in Anbetracht ihrer Vor- und Nachteile im Folgenden näher erläutern.

Betrachtet man die Idee und den Aufbau von Wort-HMM so wird schnell die Komplexität beim Erzeugen eines allumfassenden Spracherkenners deutlich. Möchte man bspw. eine komplette Sprache mit allen Wortformen, Aussprachedifferenzen und zusätzlichen regionalen Dialekten berücksichtigen, gestaltet sich die Darstellung bald als sehr unpraktikabel. Zusätzlich zum Umfang des Vokabulars müsste jedes Wort von einigen hundert bis tausend unterschiedlichen Personen aufgenommen werden, damit eine zufriedenstellende Erkennung sichergestellt werden kann. Eine solche Erstellung würde sich nicht nur als sehr zeit- sondern zusätzlich auch als sehr kostenintensiv gestalten. Kommen daher Wort-HMM in der Praxis vor, so nur zur Erkennung eines Vokabulars mit eher geringem Umfang. [1, S. 348 f.]

Um eben dieses Problem zu umgehen, eignet sich die Verwendung von Laut-HMM. Spracherkenner auf Basis phonetischer Grundelemente lassen sich einfacher und beliebiger erweitern und für das Hinzufügen neuer Wörter genügt lediglich die phonetische Umschreibung. [1, S. 363] Zwar werden bspw. bei einem Vokabular bestehend aus rund 59.000 Wörtern immer noch etwa 28.000 Triphone gezählt, jedoch müssen eben nicht mehr alle Wörter einzeln verarbeitet werden. Von den vorhanden Triphonen kommen bei obigem Beispiel außerdem mehr als die Hälfte unter zehn mal vor, sodass sich die wirklich vermehrte Nutzung nur auf einen deutlich geringeren Teil bezieht. [1, S. 353] Ein weiterer Vorteil, der durch die phonetischen Grundelemente entsteht, ist die zunehmende Sprecherunabhängigkeit bei der Erzeugung und Nutzung eines Spracherkenners auf Basis von Laut-HMM. [1, S. 363]

Im folgenden *Kapitel 5* der Seminararbeit erfolgt anhand eines Fazits ein direkte Gegenüberstellung beider Ansätze.

5 Fazit

Die Spracherkennung auf Basis statistischer Modelle und insbesondere auf Basis der HMM ist im Laufe der Jahre zu einer etablierten Technik geworden. Im Zuge der Seminararbeit wurde nach Einführung in die HMM die statistische Spracherkennung mit Hilfe dieser näher erläutert und erklärt. Grundlegend haben sich im Laufe der letzten Jahrzehnte zwei Modell-Ansätze entwickelt: zum einen die Spracherkennung mit Hilfe der Wort-HMM und zum anderen mit Hilfe der Laut-HMM.

Stellt man nun beide Ansätze gegenüber, so wird schnell klar, dass das Laut-HMM im Gegensatz um Wort-HMM eine deutlich höhere Flexibilität und über ein deutlich positiveres Kosten-Nutzen-Verhältnis verfügt. Eine mögliche Reduktion der zu einer gut funktionierenden Spracherkennung verpflichtend vorhandenen Datenmenge ist essentiell und erweist sich bei Laut-HMM deutlich größer als bei Wort-HMM.

Abschließend lässt ist auf Grundlage der Seminararbeit der Schluss ziehen, dass sich bei einem direkten Vergleich beider Ansätze die Verwendung von Laut-HMM als sinnvoller erweist.

Literaturverzeichnis

[1] B. Pfister und T. Kaufmann, Sprachverarbeitung - Grundlagen und Methoden der Sprachsynthese und Spracherkennung, Zürich: Springer, 2008.

[2] D. Jurafsky und J. H. Martin, Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Stanford, 2019.

[3] L. R. Rabiner und B. H. Juang, „An introduction to hidden Markov models," *IEEE ASSP Magazine,* Bd. 3, Nr. 1, pp. 4 - 16, 1986.