ERZEUGUNG VON LAUTSEQUENZEN FÜR UNBEKANNTE TEXTSEGMENTE MIT HILFE EINES STATISTISCHEN ANSATZES

Alexander Micheel, Hans-Günter Hirsch

Institut für Mustererkennung, Hochschule Niederrhein
alexander.micheel@hs-niederrhein.de

Abstract: Zur synthetischen Generierung eines Sprachsignals aus einem Textsegment als auch zur Definition von Referenzmodellen für Wörter im Bereich der lautbasierten Spracherkennung benötigt man ein Verfahren zur Festlegung einer Abfolge von Lauten für die Buchstabenfolge eines gegebenen Textsegments oder eines Worts, wenn das Textsegment oder das Wort nicht in einem Aussprachelexikon enthalten ist. In diesem Beitrag wird ein statistischer Ansatz zur Realisierung dieser Zuordnung einer Lautfolge vorgestellt, dessen Entwicklung durch den Bedarf zum Betrieb eines Spracherkennungsdienstes initiiert wurde. Der statistische Ansatz beruht auf der Verwendung eines Markov Modells, das eine der Anzahl von Lauten entsprechende Zahl von Zuständen besitzt. Jeder Zustand emittiert mit einer entsprechenden Wahrscheinlichkeit eine Folge von 1 bis 3 Buchstaben. Die Zuordnung einer Lautfolge zu einem Textsegment erfolgt durch die Bestimmung der Zustandsfolge, die mit größter Wahrscheinlichkeit die Buchstabenfolge des gegebenen Textsegments emittiert. Die Vorgehensweisen zur Festlegung der Modellparameter in einer Trainingsphase sowie zur Ermittlung der Zustandsfolge für ein vorgegebenes Textsegment werden detailliert erläutert. Zur quantitativen Beschreibung der Effizienz des Verfahrens werden die Ergebnisse von Simulationsexperimenten vorgestellt. Dabei wird ein Aussprachelexikon als Referenz herangezogen, das von der Carnegie Mellon Universität (CMU) zur freien Verwendung zur Verfügung gestellt wird. Es werden die Fehlerraten bei der Festlegung der Anzahl und der Abfolge der Laute im Vergleich zur lautsprachlichen Beschreibung im Lexikon ermittelt. Dabei wird zum Test des Verfahrens ein kleinerer Anteil aller Einträge im CMU Lexikon benutzt. Die verbleibenden Anteile werden zum Training der Modellparameter herangezogen. Man erhält damit beispielsweise eine Lautfehlerrate von 12,9 %, wenn man zum Training 95 % der zufällig ausgewählten Einträge und zum Test die verbleibenden 5 % verwendet.

1 Einleitung

Zur automatischen Spracherkennung werden heutzutage in der Regel Erkennungssysteme eingesetzt, die auf der Modellierung der Laute einer Sprache mit Hilfe von Hidden Markov Modellen (HMM) beruhen. Zur Erkennung von Wörtern wird dabei ein Aussprachelexikon verwendet, in dem zu jedem Wort die zugehörige Folge von Lauten definiert ist. Durch die Verkettung der Lautmodelle kann man so ein HMM für das jeweilige Wort beschreiben. In der Regel werden Triphon Modelle verwendet, die das Auftreten eines Lauts im Kontext des vorhergehenden und des nachfolgenden Lauts berücksichtigen. Damit werden deutlich bessere Erkennungsraten im Vergleich zur Verwendung von Monophonmodellen erzielt, bei denen der Kontext unberücksichtigt bleibt. Allerdings benötigt man für die Modellierung mit Triphonen eine wesentlich größere Anzahl von HMMs, so dass daraus ein erheblich höherer Rechenaufwand resultiert. Mit dem Ansatz einer Verwendung von Lautmodellen in Kombination mit einem Aussprachelexikon ergibt sich ein Problem, wenn man ein Wort modellieren möchte, das nicht in dem Lexikon enthalten ist. Dies ist typischerweise der Fall,

wenn man beispielsweise einen Eigennamen oder einen Straßennamen, dessen Aussprache nicht bekannt ist, erkennen möchte. In diesem Fall benötigt man ein Regelwerk oder ein Verfahren, mit dem man dem unbekannten Textsegment eine Folge von Lauten zuordnet. In gleicher Weise stößt man auch im Bereich der Sprachsynthese auf diese Problematik, bei der die Abbildung des zu synthetisierenden Textes auf eine Lautfolge erforderlich ist. Um die Stelle, an der ein Verfahren zur Festlegung der Lautfolge benötigt wird, in einer praktischen Anwendung zu veranschaulichen, wird in Abbildung 1 die Vorgehensweise skizziert, die wir in einem eigenen verteilten Erkennungssystem einsetzen. Das Erkennungssystem ist zur Erkennung kleinerer bis mittlerer Wortschätze mit maximal einigen Hundert Wörtern konzipiert. Zurzeit angedachte Einsatzbereiche sind die sprachbediente Steuerung von Geräten im Bereich der Hausautomation oder allgemeiner von Geräten, die man in dem mit dem Schlüsselwort „Internet of Things“ belegten Umfeld findet.

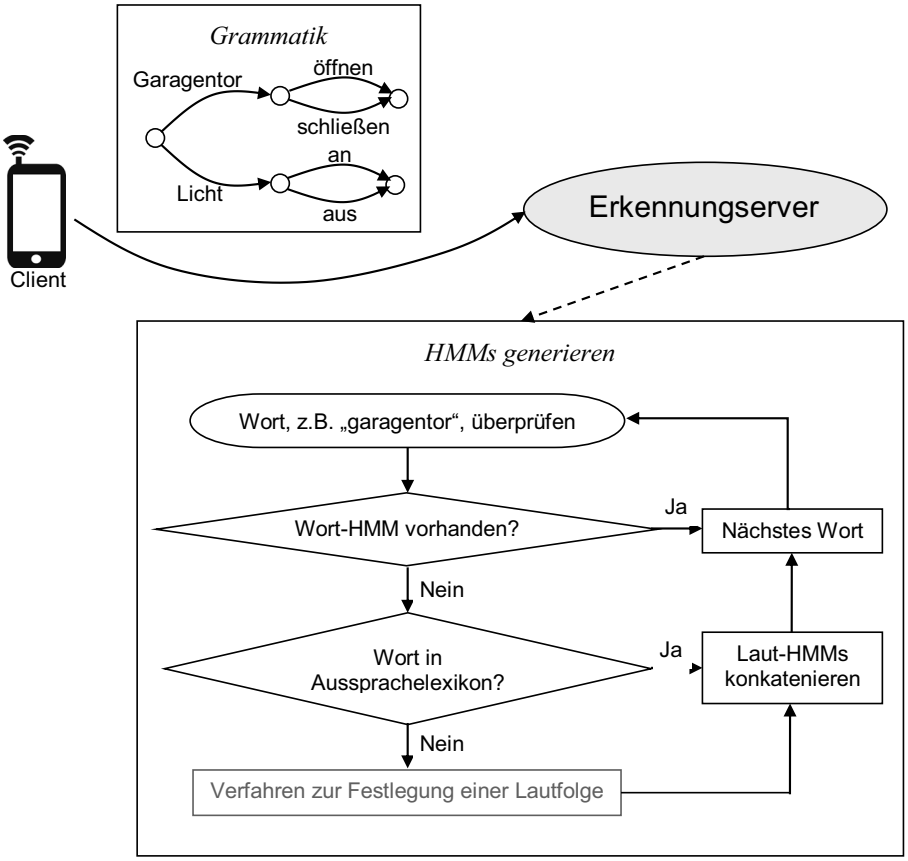

Abbildung 1: Einsatz der Lautfolgenbestimmung in einem Spracherkennungssystem

Der Erkennungsserver ist so konzipiert, dass das Clientsystem vor jedem einzelnen Erkennungsvorgang dem Serversystem optional im Rahmen eines festgelegten Protokolls eine Grammatik übermitteln kann. Damit werden unter anderem die zur Erkennung benötigten Wörter in Textform vorgegeben. Das Erkennungssystem beinhaltet eine Vielzahl von Wortmodellen für die im Bereich von Steuerungsaufgaben typischen Wörter. Diese wortbasierten HMMs wurden mit einer entsprechenden Anzahl von Äußerungen des jeweiligen Worts trainiert. Ist das entsprechende Wortmodell nicht vorhanden, so wird

versucht, mit Hilfe von Triphonmodellen und eines umfangreichen Aussprachelexikons ein entsprechendes HMM zu generieren. Findet sich das Wort nicht in dem Lexikon, wird ein Verfahren zur Festlegung der Lautfolge benötigt.

Zur phonetischen Transkription, wie man die Definition einer Lautfolge für ein gegebenes Wort oder Textsegment auch bezeichnet, existiert eine Vielzahl von Ansätzen, die man grob 2 Klassen zuordnen kann. Die erste Klasse beinhaltet regelbasierte Verfahren [1], die auf einem definierten Regelwerk der Generierung von Lautfolgen in einer Sprache beruhen. Zur Festlegung des Regelwerks werden fundierte linguistische und phonetische Kenntnisse der jeweiligen Sprache benötigt. Demgegenüber basieren die Verfahren der zweiten Klasse auf der Beschreibung der Lautgenerierung mit Hilfe eines Modells, wobei die Parameter des Modells durch die Analyse einer Vielzahl bekannter phonetischer Transkriptionen festgelegt werden. Die Modellierung beruht beispielsweise auf dem Einsatz von neuronalen Netzen [2], Markov Modellen [3], [4] oder Multigrammen [5], [6].

Im Rahmen der hier vorgestellten Untersuchungen wird ein statistischer Ansatz zur Bestimmung der Folge von Lauten vorgestellt, den man der zweiten Klasse zuordnen kann. Es wird ein Markov Modell eingesetzt, das eine der Anzahl von Monophon- bzw. Triphonmodellen entsprechende Zahl von Zuständen besitzt. Alle Übergänge zwischen beliebigen Zuständen des HMMs sind möglich. Die Emission eines Zustands besteht aus einem kurzen Textsegment, das aus bis zu 3 Buchstaben bestehen kann. Man könnte das Modell auch als Hidden Markov Modell bezeichnen, wie man es im Bereich der Spracherkennung tut. Das Attribut „versteckt“ deutet an, dass man als Ausgangswerte des Modells eigentlich nur das Textsegment erhält, ohne die dazu durchlaufene Zustands- und Lautfolge zu sehen. Das entwickelte Verfahren wird detailliert im nachfolgenden Kapitel vorgestellt. Insbesondere werden die Vorgehensweisen zum Training der Parameter des Modells und zur Festlegung der Lautfolge für ein unbekanntes Textsegment erläutert. Die Effizienz des Verfahrens wurde experimentell mit Hilfe des von der CMU veröffentlichten Aussprachelexikons [7] untersucht. Die dabei erzielten Ergebnisse werden in Form von Fehlerraten, die den Prozentsatz fehlerhafter Laute quantitativ beschreiben, in einem weiteren Kapitel dokumentiert.

2 Textgenerierung mit Hilfe eines Markov Modells

Der hier vorgestellte Ansatz beruht auf dem Einsatz eines Markov Modells. Das Modell besteht aus einer der Anzahl von Monophonen bzw. Triphonen entsprechenden Zahl von Zuständen. Konkret sind es beispielsweise 39 Zustände für das in dem CMU Aussprachelexikon verwendete Lautinventar, wenn man zunächst die Betrachtung auf die Verwendung von Monophonen beschränkt. Ergänzt wird das Modell noch um einen Startzustand und einen Endzustand, die keine Emissionen in Form von Textschnipseln erzeugen. Das Modell erlaubt den Übergang zwischen beliebigen Zuständen bis auf das Verweilen in einem Zustand und bis auf die Übergänge in den Startzustand und die Übergänge aus dem Endzustand. Start- und Endzustand dienen gemäß ihrer Bezeichnung nur als Einstiegs- und Ausstiegsknoten.

Als Parameter zur Festlegung der Übergänge wird somit eine vollständige Matrix von Übergangswahrscheinlichkeiten $p(ph_y \mid ph_x)$ benötigt, die bis auf die Wahrscheinlichkeiten $p(ph_x \mid ph_x)$ zum Verweilen in dem Zustand des Phonems ph_x und die Übergänge in den Startzustand und die Übergänge aus dem Endzustand Werte ungleich Null beinhalten können. Beim Training der Modellparameter, die durch die Analyse bekannter phonetischer Transkriptionen erfolgt, findet man eine Vielzahl von Lautübergängen, die in einer Sprache nicht auftreten und deren Übergangswahrscheinlichkeiten folglich auch den Wert Null annehmen. In jedem der Lautzustände wird als Emission ein Textschnipsel erzeugt, das aus 1

bis 3 Buchstaben besteht. In Abbildung 3 ist dazu beispielhaft die Zustandsfolge zur Generierung des Worts „house“ dargestellt.

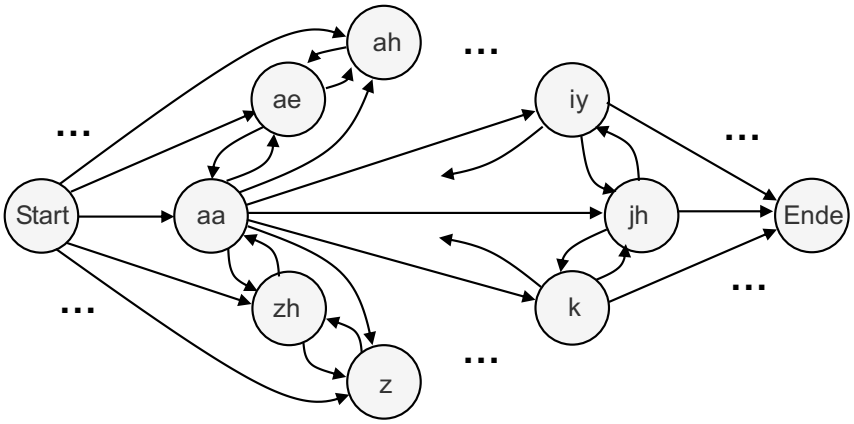

Abbildung 2: Markov Modell mit Zuständen zur Generierung der Lautfolge

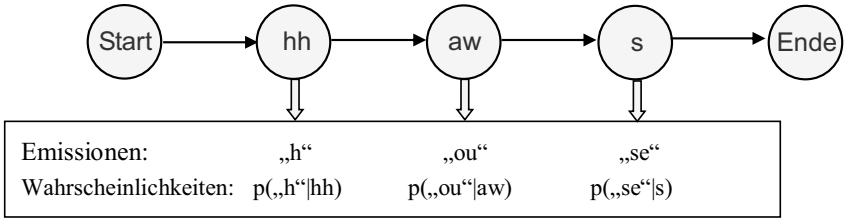

Abbildung 3: Zustandsfolge zur Generierung des Worts „house“

In dem dargestellten Beispiel erzeugt der Zustand des Lauts hh den Buchstaben „h“ mit der Emissionswahrscheinlichkeit $p(\text{"}h\text{"} \mid hh)$, der Zustand des Lauts aw die Buchstaben „ou“ mit der Wahrscheinlichkeit $p(\text{"}ou\text{"} \mid aw)$ und der Zustand des Lauts s die Buchstaben „se“ mit der Wahrscheinlichkeit $p(\text{"}se\text{"} \mid s)$. Damit lässt sich die Wahrscheinlichkeit $p(\text{"}house\text{"} \mid \{hh, aw, s\})$ der Generierung des Worts „house“ mit der Folge der Markov Zustände $\{hh, aw, s\}$ berechnen zu

$$p(\text{"}house\text{"} \mid \{hh, aw, s\}) = p(hh \mid start) \cdot p(\text{"}h\text{"} \mid hh) \cdot p(aw \mid hh) \cdot p(\text{"}ou\text{"} \mid aw) \cdot p(s \mid aw) \cdot p(\text{"}se\text{"} \mid s) \cdot p(ende \mid s)$$

Als Parameter werden für jeden Lautzustand theoretisch die Emissionswahrscheinlichkeiten $p(txt_k \mid ph_x)$ zur Generierung eines beliebigen Textschnipsels txt_k mit 1 bis 3 Buchstaben benötigt. Unter Berücksichtigung aller Permutationen, was die Aufeinanderfolge von 2 oder 3 Buchstaben angeht, erhält man dabei eine extrem große Anzahl von Emissionswahrscheinlichkeiten. In der praktischen Realisierung, was die Generierung von Textsegmenten in einer Sprache angeht, treten jedoch sehr viele Buchstabenkombinationen nicht auf und die zugehörigen Wahrscheinlichkeiten nehmen den Wert Null an. Beispielsweise erhält man im Mittel über alle Lautzustände nach einem Training der Modellparameter mit dem CMU Lexikon etwa 50 Emissionswahrscheinlichkeiten je Zustand ungleich Null. In den beiden nachfolgenden Abschnitten werden die Vorgehensweisen zur

Festlegung der Modellparameter in einer Trainingsphase und zur Bestimmung der Zustandsfolge mit größter Wahrscheinlichkeit für ein unbekanntes Textsegment erläutert.

2.1 Training

Zur Festlegung der Parameter des Markov Modells werden in einer Trainingsphase die Kenntnisse über das Auftreten von Lautfolgen und die Zuordnung von Lautfolgen zu Textsegmenten benötigt. Dazu verwendet man das Aussprachelexikon einer Sprache, wobei im Rahmen der hier vorgestellten Untersuchungen das CMU Lexikon für die englische Sprache benutzt wurde. Darin werden die Lautsequenzen für die in einer Sprache mit einer gewissen Häufigkeit verwendeten Wörtern angegeben, wie es beispielhaft in Tabelle 1 für einige Wörter dargestellt ist. Im CMU Lexikon werden etwa 130000 Textsegmente in Bezug auf die zugehörige Lautfolge definiert.

Wort	Lautfolge
needs	n-iy-d-z
panty	p-ae-n-t-iy
rocker	r-aa-k-r

Tabelle 1: Lautfolgen einiger Wörter

Zunächst lassen sich aus der Betrachtung aller Lautfolgen die Häufigkeiten und damit auch die Übergangswahrscheinlichkeiten $p(ph_y \mid ph_x)$ auf einfache Weise bestimmen, wobei dazu die Kenntnis der zugehörigen Textsegmente nicht benötigt wird. Etwas schwieriger ist die Bestimmung der Emissionswahrscheinlichkeiten $p(txt_k \mid ph_x)$. Dazu wird jedes Textsegment in einer der Anzahl von Lauten entsprechende Zahl von Buchstabenfolgen mit 1 bis 3 Buchstaben zerlegt. Dabei gibt es eine Vielzahl von Möglichkeiten der Zerlegung des Textsegments, wie es beispielhaft in Tabelle 2 für das Wort *house* dargestellt ist, dem die Lautfolge $\{hh, aw, s\}$ mit 3 Lauten zugeordnet ist.

hh	aw	s
„h“	„ous“	„e“
„h“	„ou“	„se“
„h“	„o“	„use“
„ho“	„u“	„se“
„ho“	„us“	„e“
„hou“	„s“	„e“

Tabelle 2: Mögliche Emissionen der Zustandsfolge {hh, aw, s} zur Generierung des Worts *house*

Gerade bei längeren Wörtern gibt es dabei eine sehr große Anzahl möglicher Zerlegungen des Textsegments. Aus der Verwendung aller möglichen Zerlegungen resultierte bei der praktischen Umsetzung ein sehr hoher Rechenaufwand. Daher wurden für jedes Textsegment zufällig vier Zerlegungen ausgewählt. Bei dem CMU Lexikon mit etwa 130000 Einträgen erhält man etwa 800000 mal 4 Zuordnungen von Buchstabenfolgen zu Lauten. Aus allen Zuordnungen von Buchstabenfolgen zu Lauten lassen sich die Emissionswahrscheinlichkeiten $p(txt_k \mid ph_x)$ aus den entsprechenden Häufigkeiten des Auftretens einer Zuordnung bestimmen. Da die zufällige Auswahl von 4 Zerlegungen eine deutliche Einschränkung darstellt, werden die derart bestimmten

Emissionswahrscheinlichkeiten nur als initiale Werte verwendet, die in einem weiteren Optimierungsschritt erneut berechnet werden. Dazu werden für jeden der Einträge des Lexikons die Wahrscheinlichkeiten $p(\text{"}wort\text{"} \mid \{ph_1, ph_2, ..., ph_N\})$ zur Generierung des Worts mit der zugehörigen Lautfolge in der schon zuvor beschriebenen Weise berechnet. Dabei werden alle möglichen Emissionen der Zustandsfolge $\{ph_1, ph_2, ..., ph_N\}$ betrachtet, mit denen das zugehörige Wort erzeugt werden kann. Es erfolgt die Beschränkung auf diese Zerlegung des Worts, für die die größte Wahrscheinlichkeit berechnet wird. Somit erhält man für jeden Eintrag im Lexikon nur eine Zerlegung des Textsegments in die der Anzahl der Laute entsprechende Zahl von Buchstabenfolgen. Aus dieser reduzierten Anzahl von Zuordnungen kann man erneut die Wahrscheinlichkeiten $p(txt_k \mid ph_x)$ berechnen. Den Optimierungsschritt kann man iterativ so lange wiederholen, bis man keine oder nur noch eine geringfügige Verbesserung der mittleren Gesamtwahrscheinlichkeit in einem Schritt erzielt.

2.2 Festlegung der Lautfolge für ein unbekanntes Textsegment

Im Folgenden wird die Vorgehensweise erläutert, wie man mit dem Markov Modell, dessen Parameter in der im vorherigen Abschnitt beschriebenen Form bestimmt wurden, die Folge von Lauten für ein Textsegment, das nicht in dem Lexikon enthalten ist, festlegt. Ausgehend vom Startzustand betrachtet man alle Lautzustände, die die Emission der ersten 1 bis 3 Buchstaben des zu analysierenden Textsegments erlauben. Dazu kann man die zugehörigen Wahrscheinlichkeiten $p(ph_1 \mid start) \cdot p(\text{"}xx_1\text{"} \mid ph_1)$ zur Generierung eines ersten Teilsegments xx_1 mit einem Lautzustand ph_1 berechnen. Ausgehend von allen Zuständen, die ein erstes Teilsegment emittieren können, wird der Übergang zu jeweils allen anderen Zuständen betrachtet, die ein darauf folgendes Teilsegment mit 1 bis 3 Buchstaben generieren können. Für die dabei insgesamt auftretenden 2 Übergänge und 2 Emissionen werden für alle Möglichkeiten der Kombination zweier Laute wiederum die Wahrscheinlichkeiten $p(ph_1 \mid start) \cdot p(\text{"}xx_1\text{"} \mid ph_1) \cdot p(ph_2 \mid ph_1) \cdot p(\text{"}xx_2\text{"} \mid ph_2)$

berechnet. Da es an dieser Stelle eine sehr hohe Anzahl von Zustandskombinationen gibt, die ein erstes und ein zweites Teilsegment des Textsegments emittieren, erfolgt an dieser Stelle die Beschränkung auf die 10000 Kombinationen mit der höchsten Wahrscheinlichkeit. Die bis zu dieser Stelle emittierten Segmente können theoretisch aus 2 bis 6 Buchstaben bestehen.

Die für das zweite Teilsegment angewandte Vorgehensweise wird für das dritte und eventuell alle weiteren Teilsegmente wiederholt. Dabei werden immer wieder alle möglichen Zustandsübergänge und Zustände zur Generierung eines weiteren Teilsegments des Textsegments betrachtet. Nach jedem Optimierungsschritt erfolgt die Beschränkung auf die 10000 wahrscheinlichsten Zustandsfolgen. Erreicht man bei einem Optimierungsschritt die Generierung des vollständigen Textsegments, so wird zur Berechnung der Gesamtwahrscheinlichkeit nur noch der Übergang in den Endzustand berücksichtigt. Die Festlegung der Lautfolge erfolgt abschließend aus der Ermittlung der Zustandsfolge mit der größten Gesamtwahrscheinlichkeit.

$$\underset{alle\ Folgen}{MAX} \sqrt[2 \cdot Nz+1]{p(ph_1 \mid start) \cdot p(\text{"}xx_1\text{"} \mid ph_1) \cdot p(ende \mid ph_{Nz}) \cdot \prod_{j=2}^{Nz} \left(p(\text{"}xx_j\text{"} \mid ph_j) \cdot p(ph_j \mid ph_{j-1}) \right)}$$

Der Wurzelterm wird dabei individuell für jede der Zustandsfolgen berechnet, mit denen das Textsegment korrekt erzeugt werden kann. Nz ist die Anzahl der Zustände der jeweiligen Zustandsfolge. Durch die Wurzelbildung erfolgt eine wertemäßige Skalierung auf die $(2 \cdot Nz + 1)$ Multiplikationsterme, deren Anzahl bei unterschiedlichen Zustandsfolgen

verschieden sein kann. Die Abhängigkeit von der jeweiligen Zustandsfolge bleibt in der mathematischen Darstellung unberücksichtigt.

3 Experimentelle Untersuchungen

Um quantitative Angaben zur Güte des eingesetzten Modells bei der Generierung von Lautfolgen machen zu können, wird das Aussprachelexikon herangezogen, das Entwicklern und Forschenden von der CMU zur freien Verwendbarkeit zur Verfügung gestellt wird [7]. Das Lexikon enthält insgesamt etwa 130000 Einträge in englischer Sprache, die die jeweilige Lautfolge zur Festlegung der englischen Aussprache im amerikanischen Bereich definieren. In dem Lexikon sind etwa 8900 Einträge vorhanden, die eine alternative Aussprachevariante von bereits im Lexikon enthaltenen Wörtern oder Textsegmenten beinhalten. Das Lexikon beruht auf der Verwendung eines Inventars von 39 Lauten, die in Tabelle 3 in einer ASCII Notation aufgelistet sind.

aa	ae	ah	ao	aw	ay	b	ch	d	dh	eh	er	ey
f	g	hh	ih	iy	jh	k	l	m	n	ng	ow	oy
p	r	s	sh	t	th	uh	uw	v	w	y	z	zh

Tabelle 3: Lautinventar des CMU Aussprachelexikons

Zur Durchführung der Experimente werden alle Einträge des CMU Lexikons 2 Gruppen zugeordnet. Mit den Einträgen der ersten Gruppe, die einen wesentlich größeren Anteil aller Einträge beinhaltet, wird ein Training der Parameter des Markov Modells in der Weise vorgenommen, die im vorhergehenden Kapitel erläutert wurde. Es werden die drei in Tabelle 4 aufgeführten Zuordnungen als experimentelle Varianten untersucht.

Variante	Trainingsdaten	Testdaten
90_10	90%	10%
95_5	95%	5%
100_5	100%	5%

Tabelle 4: Prozentuale Aufteilung des CMU Lexikons für Training und Test

Als quantitative Angabe zur Beurteilung der Güte des Verfahrens wird eine Phonem-Fehlerrate definiert, die die Anzahl von Verwechslungen, von Auslöschungen und von Einfügungen beinhaltet:

$$Phonemfehlerrate = \frac{Verwechslungen + Auslöschungen + Einfügungen}{Gesamtzahl\ der\ Phoneme}$$

Eine oder mehrere Auslöschungen treten auf, wenn die Anzahl der mit dem Verfahren generierten Laute kleiner als die Anzahl im Aussprachelexikon ist. Umgekehrt kommt es zu einer oder mehreren Einfügungen, wenn die Anzahl der generierten Laute größer als die im Lexikon ist. Die Anzahl aller Laute, die bei den Einträgen auftreten, die zum Test verwendet werden, definiert die Gesamtzahl der Phoneme.

Bei der Verwendung eines Markov Modells mit 39 Zuständen zur Beschreibung der in Tabelle 3 enthaltenen 39 Monophone stellen sich für die drei experimentellen Varianten die in Tabelle 5 aufgeführten Phonem-Fehlerraten ein. Alternativ wird ein Modell mit der wesentlich größeren Zahl von 19000 Zuständen verwendet, in dem man die Generierung eines Lauts als Triphon in Abhängigkeit des vorhergehenden und des nachfolgenden Lauts betrachtet. Dabei stellen sich die in Tabelle 6 aufgeführten Fehlerraten ein.

Experiment	Phonem-Fehlerrate
Mono_90_10	28,50 %
Mono_95_5	28,46 %
Mono_100_5	28,41 %

Tabelle 5: Phonem-Fehlerraten bei Einsatz von Monophonmodellen

Experiment	Phonem-Fehlerrate
Tri_90_10	13,26 %
Tri_95_5	12,92 %
Tri_100_5	12,80 %

Tabelle 6: Phonem-Fehlerraten bei Einsatz von Triphonmodellen

Bei Verwendung der wesentlich größeren Zahl von Triphonzuständen stellen sich dabei deutlich geringere Fehlerraten ein. Mit der großen Anzahl von Zuständen wird eine wesentlich differenziertere Generierung von Buchstabenkombinationen ermöglicht.

4 Zusammenfassung

Es wurde ein statistischer Ansatz vorgestellt, um die phonetische Transkription eines Textsegments, dessen zugehörige Lautfolge unbekannt ist, vorzunehmen. Auf Grund der gewählten Struktur des Markov Modells, bei der der Übergang von einem Lautzustand in jeden anderen möglich ist, kann der im Bereich der Spracherkennung eingesetzte Viterbi Algorithmus nur schwerlich angewendet werden. Der Rechenaufwand für das entwickelte Decodierungsverfahren ist daher relativ hoch, wobei schon eine schrittweise Beschränkung auf die wahrscheinlichsten Zustandsfolgen stattfindet. Die erzielten Phonem-Fehlerraten reichen zurzeit noch nicht an die besten Ergebnisse heran, die mit alternativen Ansätzen erzielt wurden. An verschiedenen Stellen besteht jedoch noch ein deutliches Verbesserungspotential. Beispielsweise wurde der spezielle Fall, dass der Buchstabe „x“ häufig die Generierung der beiden aufeinander folgenden Laute „k“ und „s“ benötigt, bisher nicht in dem Modell erfasst.

Literatur

[1] Stock, P.: P-TRA- Eine Programmiersprache zur phonetischen Transkription, in: W. Hesss; W. Sandlemeier Eds. “Beiträge zur Angewandten und Experimentellen Phonetik“, Franz Steiner Verlag, Stuttgart, 1992

[2] Adamson, M. and R. Damper, R.: A Recurrent Network that Learns to Pronounce English Text, in Proceedings of the International Conference on Spoken Language Processing, Philadelphia, USA, Oct. 1996

[3] Kantor, A., Hasegawa-Johnson, M.: HMM-based Pronunciation Dictionary Generation. University of Illinois at Urbana-Champaign, USA, 2011

[4] Taylor, P.: Hidden Markov Models for Grapheme to Phoneme Conversion, In Proc. of the Interspeech conference, pp. 1973–1976, Portugal, Sep. 2005

[5] Fisher, W.M.: A statistical text-to-phone function using ngrams and rules, National Institute of Standards and Technology, in Proc. of ICASSP, pp. 649-652, 1999

[6] Deligne, S., Yvon, F., and Bimbot, F.: Variable length sequence matching for phonetic transcription using joint multigrams, in Proc. Eurospeech, Madrid, Sep. 1995

[7] “The CMU pronouncing dictionary”, http://www.speech.cs.cmu.edu/cgi-bin/cmudict