Rheinisch Westfälische Technische Hochschule Aachen
Lehrstuhl für Informatik VI
Prof. Dr.-Ing. Hermann Ney

Speech Recognition and Language Processing

Confidence Measures in Speech Recognition

(Konfidenzmaße in der Spracherkennung)

Daniel H. Peger

Matrikelnummer 222 886

31. Juli 2003

Betreuer: Dr. Ralf Schlüter

Inhaltsverzeichnis

1 Einleitung **6**
1.1 Hypothesendarstellung 8

2 Konfidenzmaße **9**
2.1 Wortposition 9
2.1.1 Wortposition in Wortgraphen 9
2.1.2 Wortposition N-Besten Listen 11
2.2 Heuristische Konfidenzmaße 12
2.2.1 Acoustic Stability 12
2.2.2 Hypothesis Density 13
2.3 Statistisch motivierte Konfidenzmaße 14
2.3.1 Phonembasiert 14
2.3.2 Word Posterior Wahrscheinlichkeiten 16

3 Berechnung von Word Posterior Wahrscheinlichkeiten **18**
3.1 Wortgraphen basierte Definition 18
3.2 N-Besten Listen basierte Definition 23

4 Vergleich verschiedener Verfahren **24**
4.1 Datensätze 25
4.2 Evaluierungsmaße 26
4.3 Resultate 27

5 Anwendung von Konfidenzmaßen in der Praxis **29**

6 Zusammenfassung und Ausblick **31**

Literaturverzeichnis **31**

Tabellenverzeichnis

1 Charakterisierung der Testdatensätze 25
2 CERs für $C_{D_{avg}}$ und C_A 26
3 CERs für die Verfahren auf N-Besten Listen und Wortgraphen 26
4 CERs für die Verfahren auf N-Besten Listen und Wortgraphen 27

Abbildungsverzeichnis

1 Architektur eines Spracherkennungssystems für kontinuierliche Sprache . . . 6
2 Einfacher Wortgraph 8
3 Beispiel für die Wortzuordnung durch den Levensthein Algorithmus 11
4 Hypothesendichte zu verschiedenen Zeitpunkten in einem Wortgraphen . . . 14
5 Hidden Markov Modell mit 3 alternativen Hypothesen 15
6 Wortgraph mit zu bewertender Worthypothese 16
7 Veränderung der CER durch den akustischen Skalierungsfaktor 23
8 Variation des Wortgraphen aus Abbildung 2 24
9 DET-Kurven auf dem ARISE- und Verbmobil-Datensatz 28

10 DET-Kurven auf dem NAB 20k-, dem NAB 64k und dem Broadcast News-Datensatz 29
11 Beispiel Dialog zur expliziten Verifikation 30
12 Beispiel Dialog zur impliziten Verifikation 31

1 Einleitung

Spracherkennung und insbesondere die damit verbundene Steuerung von Programmen oder allgemeinen Abläufen über Sprache spielen in der heutigen Gesellschaft eine immer größere Rolle. Unabhängig davon für welche Anwendung die Erkennung gesprochener Sprache verwendet wird, sind die Grundlagen eines Erkennungsprozesses doch immer die gleichen oder sind sich zumindest sehr ähnlich.

Im groben können 2 verschiedene Ansätze in der Spracherkennung unterschieden werden. *Einzelworterkenner* sind die einfachsten Spracherkenner und können zur Bedienung oder Steuerung von Geräten und Systemen eingesetzt werden, bei denen beispielsweise die Hände des Bedieners für andere Aufgaben freigehalten werden sollen. Für komplexere Anwendungen wie Diktiersysteme oder Auskunftsdialoge ist die Erkennung *kontinuierlicher Sprache* erforderlich. Diese beiden Erkennungsaufgaben unterscheiden sich dadurch, dass bei der Einzelworterkennung die Start- und Endzeiten der einzelnen Wörter dem System bekannt sind. Bei älteren Diktiersystemen wurde dies zum Beispiel durch explizite Pausen nach jedem gesprochenen Wort erreicht. Bei der Erkennung kontinuierlicher Sprache sind sowohl die zeitlichen Grenzen der Wörter als auch die Anzahl der Wörter in einer zu erkennenden Äußerung nicht bekannt.

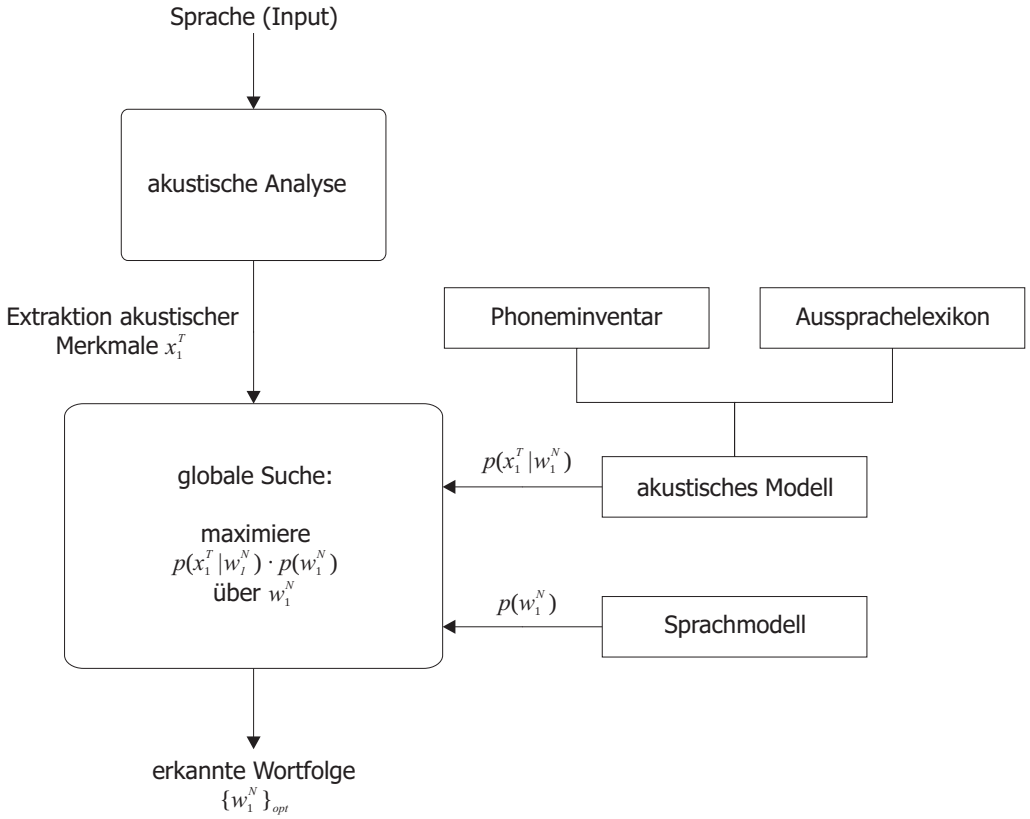


Abbildung 1: Architektur eines Spracherkennungssystems für kontinuierliche Sprache nach [8]. Das System beinhaltet eine Phonemanalyse, Ausspracheregeln und ein Sprachmodell. Wobei $x_1^T = x_1, \ldots, x_T$ eine Folge von T akustischen Beobachtungen und $w_1^N = w_1, \ldots, w_N$ eine Folge von N Wörtern ist, mit $1 \leq N \leq T$.

Abbildung 1 veranschaulicht den generellen Ablauf eines Spracherkennungsprozesses für die Erkennung kontinuierlicher Sprache. Aus dem analogen Eingangssignal werden *akustische Merkmale* x_1^T extrahiert. Mittels des akustischen Modells $p(x_1^T|w_1^N)$ und des Sprachmodells $p(w_1^N)$ wird nun eine globale Suche über der *Bayes'schen Entscheidungsregel* ausgeführt, um die beste Wortfolge $\{w_1^N\}$ zu finden. Hierzu muss die Wahrscheinlichkeit für eine Wortfolge gegeben die akustische Beobachtung $p(w_1^N|x_1^T)$ maximiert werden.

$$\begin{aligned}\{w_1^N\}_{opt} &= \underset{w_1^N}{\operatorname{argmax}} \left\{p(w_1^N|x_1^T)\right\} \\ &= \underset{w_1^N}{\operatorname{argmax}} \left\{\frac{p(x_1^T|w_1^N)^\alpha \cdot p(w_1^N)^\beta}{p(x_1^T)}\right\} \\ &= \underset{w_1^N}{\operatorname{argmax}} \left\{p(x_1^T|w_1^N)^\alpha \cdot p(w_1^N)^\beta\right\}. \end{aligned} \tag{1}$$

Die beiden Faktoren α und β sind sogenannte Skalierungsfaktoren. Mit diesen Faktoren werden das akustische und das Sprachmodell skaliert. In der Praxis liegt der Wert für α (den Skalierungsfaktor des akustischen Modells) bei ungefähr 1 und β der Faktor für das Sprachmodell liegt zwischen 10 und 25 [1]. Die konkreten Werte müssen für jedes Erkennungssystem neu ermittelt werden. Dazu werden Trainingsdaten verwendet und die Skalierungsfaktoren so gewählt, dass man die beste Erkennungsleistung erhält.

Die Wahrscheinlichkeit für die gemachte akustische Beobachtung $p(x_1^T)$ wird meistens bei der Berechnung von $\{w_1^N\}_{opt}$ weggelassen, da dieser Term unabhängig von der jeweils betrachteten Wortfolge ist und somit keinen Einfluss auf die Maximierung hat.

Wie in Abbildung 1 und Gleichung (1) zu erkennen ist, bestimmen aktuelle Spracherkenner zu einem akustischen Signal die wahrscheinlichste Wortfolge. Nun ist für viele Anwendungen die korrekte Erkennung der gesamten Wortfolge nur zweitrangig, da es bei diesen Anwendungen nur um bestimmte Schlüsselwörter eines Satzes und den damit zusammenhängenden Kontext geht. In diesem Zusammenhang treten zwei neue Probleme auf. Zum ersten gilt es die Schlüsselwörter in einer Wortfolge zu identifizieren und zum zweiten wird eine Art Klassifizierer benötigt, der für ein einzelnes Wort eines erkannten Satzes entscheidet ob dieses Wort wahrscheinlich korrekt erkannt wurde oder nicht. Diese Ausarbeitung beschäftigt sich im Folgenden mit dem letzteren Problem und genauer mit sogenannten *Konfidenzmaßen* (eng.: *Confidence Measures*), die zu einem gegebenen Wort innerhalb eines Satzes eine Bewertung darstellen, anhand derer ein einfacher Klassifizierer die Einstufung in *„richtig“* oder *„falsch“* vornehmen kann.

Nachdem in den Abschnitten 1.1 und 2.1 zunächst die für die weitere Arbeit grundlegenden Begriffe erläutert wurden, werden in Abschnitt 2 verschiedene Ansätze zur Konfidenzbewertung einzelner Wörter vorgestellt. Die hier betrachteten Verfahren beruhen hauptsächlich auf [3, 6, 11, 12, 19]. In Abschnitt 3 wird näher auf einen relativ neuen Ansatz zur Wortbewertung eingegangen, der auf [14, 17] beruht. Die vorgestellten Verfahren werden in Abschnitt 4 miteinander verglichen. Ein Beispiel für die Anwendung von Konfidenzmaßen in der Praxis wird in Abschnitt 5 gegeben. Abschließend wird in Abschnitt 6 nochmals ein kurzer Überblick über die hier gemachten Beobachtungen und Schlussfolgerungen gegeben, der mit einem Ausblick auf zukünftige Entwicklungen auf dem Gebiet der Konfidenzmaße endet.

[1] Die stärkere Gewichtung des Sprachmodells gegenüber dem akustischen Modell ist auf die höhere Verlässlichkeit des Sprachmodells zurückzuführen.

1.1 Hypothesendarstellung

Bevor im Weiteren auf die verschiedenen Verfahren zur Konfidenzbestimmung eingegangen wird, müssen noch die beiden grundlegenden Darstellungsweisen von Satzhypothesen vorgestellt werden. Diese spielen eine wichtige Rolle für die Umsetzung der unterschiedlichen Konfidenzmaße. Um die vom Erkennungssystem ermittelten Satzhypothesen für die Weiterverarbeitung (z.B. zur Konfidenzbewertung) zwischenzuspeichern, werden zur Zeit entweder N-Besten Listen oder Wortgraphen verwendet.

N-Besten Listen Bei N-Besten Listen handelt es sich um simple Listen mit 100 – 1000 Einträgen von ermittelten Satzhypothesen, die üblicherweise ihrer a-posteriori oder Verbundwahrscheinlichkeit nach absteigend geordnet sind. Die erste und somit wahrscheinlichste Satzhypothese in einer solchen Liste ist die vom Erkennungssystem ermittelte beste Wortfolge $\{w_1^N\}_{opt}$.

Wortgraphen Wortgraphen sind gerichtete, a-zyklische und gewichtete Graphen mit einem dedizierten Anfangs- und Endknoten. Eine Kante eines Wortgraphen von einem Knoten an Position τ zu einem Knoten an Position t stellt eine Worthypothese $[w; \tau, t]$ dar und ist mit der akustischen Wahrscheinlichkeit für diese Hypothese gewichtet.

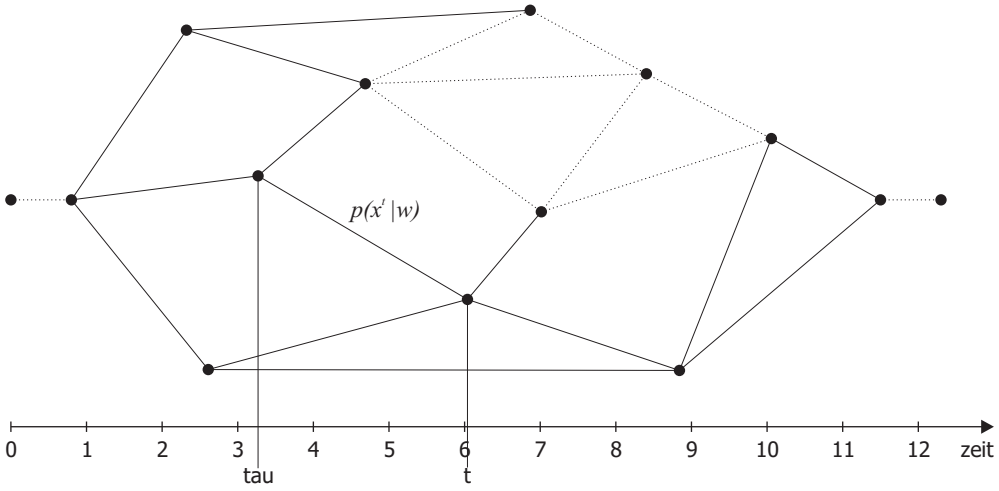


Abbildung 2: Einfacher Wortgraph. Die gestrichelten Linien stellen *Pausenkanten* (eng.: *Silence Edges*) dar. Die akustischen Wahrscheinlichkeiten für die einzelnen Kanten wurden in diesem Beispiel vernachlässigt.

Wortgraphen stellen aufgrund ihres geringen Speicherplatzbedarfs im Vergleich zu einer N-Besten Liste gleicher Komplexität[2] die zur Zeit favorisierte Darstellungsweise für Satzhypothesen dar. Desweiteren ist es leicht möglich aus einem Wortgraphen die entsprechende N-Besten Liste zu rekonstruieren, umgekehrt ist dieses nicht möglich, da die Informationen für die Anfangs- und Endzeiten einer Worthypothese bei Darstellung als N-Besten Liste nicht erhalten bleiben.

Auf Details in der Darstellungsweise, die sich bei den verschiedenen Implementierungen der Verfahren unterscheiden, wird in den entsprechenden Abschnitten eingegangen.

[2] Gleicher Anzahl an dargestellten Satzhypothesen.

2 Konfidenzmaße

Wie bereits in Abschnitt 1 angesprochen sollen Konfidenzmaße für einzelne Wörter innerhalb eines erkannten Satzes bestimmen ob diese Wörter korrekt erkannt wurden. Wurde beispielsweise der Satz „Markus sitzt alleine in seinem Auto.“ durch das verwendete Spracherkennungssystem als „Neun Markus sitzt ins einem Auto.“ erkannt, müsste die Ausgabe eines optimalen Konfidenzmaßes angewendet auf jedes einzelne der Wörter (f, r, r, f, f, r) lauten wobei f für einen Wortfehler bei der Erkennung (*Insertion*, *Substitution*, *Deletion*) steht und r für ein richtig erkanntes Wort. Wird ein vom Spracherkennungssystem richtig erkanntes Wort durch das Konfidenzmaß als „falsch“ eingestuft oder umgekehrt, spricht man von einem *Taggingfehler* (eng.: *Tagging Error*). Diese Taggingfehler spielen in Abschnitt 4.2 eine zentrale Rolle, wenn es um die Evaluierung der verschiedenen Konfidenzmaße geht.

2.1 Wortposition

Bei den Verfahren zur Ermittlung von Konfidenzmaßen ist – wie bereits in Abschnitt 1 – grundsätzlich zwischen solchen zu unterscheiden, die auf kontinuierlicher Sprache arbeiten, und solchen, die auf die Erkennung einzelner Wörter ausgelegt sind. Da bei der Einzelworterkennung in der Regel[3] keine Information über einen Kontext, in dem die Wörter stehen, bekannt ist, entsprechen Konfidenzmaße für diese Art der Spracherkennung weitestgehend der Wahrscheinlichkeit mit der das Wort erkannt wurde. Wobei natürlich auch für diesen Zweck diverse Verfeinerungen und Alternativen entwickelt wurden. Als Konfidenzmaß für die Einzelworterkennung kann beispielsweise die *Likelihood Ratio* verwendet werden. Hier wird die Wahrscheinlichkeit der besten Worthypothese in Verhältnis zu den Wahrscheinlichkeiten der alternativen Hypothesen gesetzt. Ist dieser Wert sehr hoch, so wurde das Wort mit hoher Wahrscheinlichkeit korrekt erkannt. Wenn der Wert allerdings gegen 1 geht, bedeutet dies, dass die Wahrscheinlichkeit für das erkannte Wort nur geringfügig größer ist als die der Alternativhypothesen und somit bestünde in diesem Fall eine relativ große Unsicherheit bezüglich des betrachteten Wortes.

Im weiteren werden ausschließlich Verfahren betrachtet, welche eine Konfidenz für einzelne Worte eines kontinuierlichen Sprachsignals zu bestimmen versuchen. Bei diesen Verfahren kann und wird der Kontext eines Wortes innerhalb einer Äußerung berücksichtigt, um ein möglichst gutes Erkennungsergebnis zu erhalten. Der Kontext eines Wortes ist – formal gesehen – durch seine Position und somit durch seine Vorgänger- und Nachfolgerwörter definiert. Das Problem, welches allerdings mit dieser Definition einhergeht, ist, dass – abhängig von der gewählten Darstellungsweise der Satzhypothesen – die Position einer Worthypothese nicht eindeutig definiert ist.

2.1.1 Wortposition in Wortgraphen

Werden die Hypothesen mittels eines Wortgraphen dargestellt, so ist die absolute Position der einzelnen Worthypothesen innerhalb der verschiedenen Satzhypothesen unbekannt. Das kommt daher, weil verschiedene Satzhypothesen, welche ein und die selbe Worthypothese an jeweils verschiedenen Positionen enthalten, im Wortgraphen durch die gleiche

[3] Ausnahmen von dieser Annahme bilden zum Beispiel Spracherkennungssysteme, die vom Benutzer eine Spracheingabe mit Pausen zwischen den einzelnen Wörtern verlangen, aber dennoch zusammenhängende Texte erkennen können.

Wortkante laufen können. Um dennoch eine Aussage über die Position der Worthypothesen innerhalb einer Satzhypothese machen zu können, werden die Satzhypothesen mit zeitlichen Einschränkungen versehen.

Bei der Betrachtung von Wortgraphen ist eine Satzhypothese somit nicht nur als bloße Folge von Wörtern definiert, sondern als Folge von Wörtern mit festen Zeiten. $[w;\tau,t]_1^N = [w_1;\tau_1,t_1],[w_2;\tau_2,t_2],\ldots,[w_N;\tau_N,t_N]$ ist eine Satzhypothese, bestehend aus N Worthypothesen. Jede einzelne Worthypothese ist, wie bereits kurz in Abschnitt 1.1 erwähnt, mit einer Anfangszeit τ und einer Endzeit t versehen. Es gilt hierbei, dass $\tau_1 = 1$ ist und $t_N = T$ ist[4]. Desweiteren ist die Anfangszeit eines Wortes w_n durch die Endzeit des Vorgängerwortes w_{n-1} definiert oder genauer $\tau_n = t_{n-1} + 1$.

Die bereits in Abschnitt 1 vorgestellte Bayes'sche Entscheidungsregel zur Berechnung der optimalen Wortfolge $\{w_1^N\}_{opt}$ lässt sich nun mittels den oben definierten Satz- und Worthypothesen folgendermaßen darstellen

$$\begin{aligned}
\{[w;\tau,t]_1^N\}_{opt} &= \underset{[w;\tau,t]_1^N}{\operatorname{argmax}} \left\{p([w;\tau,t]_1^N|x_1^T)\right\} \\
&= \underset{[w;\tau,t]_1^N}{\operatorname{argmax}} \left\{\frac{p(x_1^T|[w;\tau,t]_1^N)^\alpha \cdot p(w_1^N)^\beta}{p(x_1^T)}\right\} \\
&\overset{(Viterbi\ Approximation)}{=} \underset{[w;\tau,t]_1^N}{\operatorname{argmax}} \left\{\frac{\prod_{n=1}^N (p(x_{\tau_n}^{t_n}|w_n)^\alpha \cdot p(w_n|w_1^{n-1})^\beta)}{p(x_1^T)}\right\}. \qquad (2)
\end{aligned}$$

Der letzte Schritt von Gleichung (2) bedarf noch der exakten Definitionen des Sprach- und des akustischen Modells

Sprachmodell:

$$p(w_1^N) = \prod_{n=1}^N p(w_n|w_1^{n-1}). \qquad (3)$$

Akustisches Modell (Viterbi Approximation):

$$\begin{aligned}
p(x_1^T|[w;\tau,t]_1^N) &= \prod_{t=1}^T p(x_t|s_t)\cdot p(s_t|s_1^{t-1}) \\
&= \prod_{n=1}^N \prod_{t=\tau_n}^{t_n} p(x_t|s_t)\cdot p(s_t|s_1^{t-1}) \\
&= \prod_{n=1}^N p(x_{\tau_n}^{t_n}|w_n). \qquad (4)
\end{aligned}$$

$s_1^T = s_1, s_2, \ldots, s_T$ ist dabei die vom Spracherkennungssystem ermittelte, beste Zustandsfolge des HMM (s. Abbildung 5) für die akustische Beobachtung.

In der Praxis wird das Sprachmodell verwendet, um w_1^{n-1} also die Vorgängerwörter – auch *Historie* von w_n – zu bewerten. Ein m-gram Sprachmodell gibt somit die Wahrscheinlichkeit für ein Wort an unter Berücksichtigung von maximal $(m-1)$ Vorgängerwörtern. Es ergibt sich somit mit einem m-gram Sprachmodell $p(w_n|w_1^{n-1}) = p(w_n|h_1^{m-1})$, wobei $h_1^{m-1} = w_1, w_2, \ldots, w_{i-1}$ mit $w_j = \emptyset, \forall j \leq 0$ die maximal $(m-1)$ Vorgänger von Wort w_i sind. Diese Vorgänger werden im folgenden auch mit $h_1^{m-1} = h_1, h_2, \ldots, h_{m-1}$ bezeichnet.

[4] T ist die Länge der akustischen Beobachtung x_1^T.

2.1.2 Wortposition N-Besten Listen

Das zu lösende Problem bei der Hypothesendarstellung durch N-Besten Listen unterscheidet sich grundlegend von dem eben betrachteten Problem mit Wortgraphen. Im Gegensatz zu Wortgraphen ist die absolute Position der Wörter in den Satzhypothesen einer N-Besten Liste bekannt. Allerdings kann sich die Position einer Worthypothese in den verschiedenen Satzhypothesen unterscheiden. Dies kommt durch Fehler, die während des Erkennungsprozesses gemacht wurden.

Im Allgemeinen werden 3 verschiedene Arten von Erkennungsfehlern unterschieden. Es gibt *Substitutions*, *Insertions* und *Deletions*. Bei einer *Substitution* wurde ein Wort des gesprochenen Satzes mit einem anderen vertauscht, eine *Insertion* ist das Einfügen eines Wortes, das im ursprünglichen Satz nicht vorkam, und eine *Deletion* ist das Löschen eines Wortes aus dem gesprochenen Satz. Abbildung 3 verdeutlicht die Auswirkungen dieser Erkennungsfehler auf die Position einer Worthypothese in 2 verschiedenen Satzhypothesen.

Um trotz dieser Erkennungsfehler ein verlässliche Zuordnung von 2 Worthypothesen auf Basis der absoluten Wortposition vornehmen zu können, kann der Levensthein Algorithmus [7] angewandt werden. Dieses Verfahren erhält 2 Sätze und eine Wortposition als Eingabe und ermittelt für das Wort an der übergebenen Position aus dem ersten Satz das Wort aus dem zweiten Satz, welches diesem am ehesten entspricht. $\mathcal{L}_m(w_1^N, v_1^{N'}) = v$ würde somit das Wort v aus $v_1^{N'}$ ermitteln, welches w_m entspricht. Die Zuordnung der Wörter wird so vorgenommen, dass die Fehlerzahl bei Zuordnung des gesamten Satzes minimal ist. Abbildung 3 verdeutlicht diese Anpassung anhand eines einfachen Beispiels.

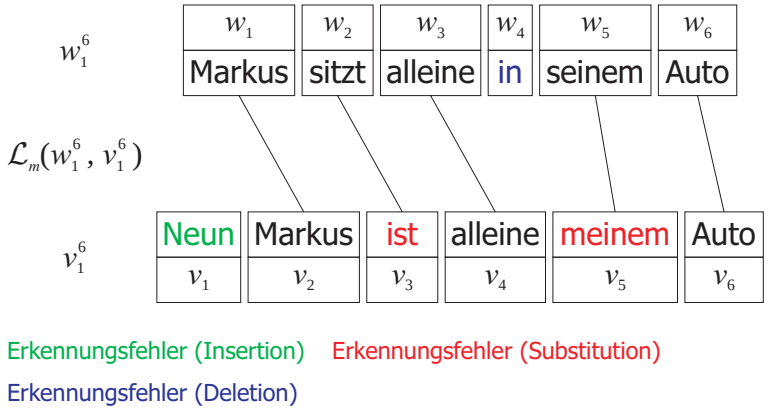


Abbildung 3: Beispiel einer Wortzuordnung durch den Levensthein Algorithmus. In der Abbildung werden die Auswirkungen der verschiedenen möglichen Fehler verdeutlicht. Eine *Substitution* ändert nichts an der absoluten Position der Wörter. *Insertions* und *Deletion* hingegen verschieben die Wörter um jeweils eine Position nach vorne bzw. nach hinten.

Durch die Verwendung des Levensthein Algorithmus zur Satzanpassung müssen die Satzhypothesen nicht, wie bei Wortgraphen, mit zeitlichen Einschränkungen versehen werden. So müssen unter Verwendung von N-Besten Listen auch keine besonderen Betrachtungen von verschiedenen Worthypothesen gemacht werden, wie sie in Abschnitt 3.1 beschrieben sind.

Die im folgenden vorgestellten Verfahren dienen alle der Verarbeitung kontinuierlicher Sprache. Diese Verfahren lassen sich grob anhand ihrer Funktionsweise in *heuristische*

Konfidenzmaße (s. Abschnitt 2.2) und *statistische Konfidenzmaße* (s. Abschnitt 2.3) aufteilen. Heuristische Verfahren verwenden Informationen, die durch empirische Methoden aus dem Erkennungsprozess als solchem gewonnen werden, wohingegen bei den statistischen Verfahren neue Modelle auf statistischer Basis entwickelt werden.

2.2 Heuristische Konfidenzmaße

Zwei der wohl bekanntesten heuristischen Konfidenzmaße sind die *Acoustic Stability* und die *Hypothesis Density* (de.: *Hypothesendichte*). Neben diesen beiden „Hauptverfahren" existieren noch diverse andere Konfidenzmaße. Alleine in [12] werden insgesamt 19 unterschiedliche Kriterien vorgestellt, die in einen Merkmalsvektor eingetragen und dann mittels klassischer Dimensionsreduzierender Verfahren zu einem linearen Klassifizierer zusammengefasst werden. In [12] wird ebenfalls noch ein neuronales Netz mit den verschiedenen Konfidenzmaßen trainiert und anschließend zur Klassifikation genutzt. Die Verwendung von mehreren parallelen Ansätzen ist bei heuristischen Konfidenzmaßen durchaus üblich und wird ebenfalls wie in [3] zur Kombination von statistischen und heuristischen Konfidenzmaßen verwendet. Hier wird ein statistisches Konfidenzmaß, das auf den akustischen Wahrscheinlichkeiten in *Viterbi Approximation*[5] beruht, mit 3 heuristischen Verfahren, einer Variation der Hypothesis Density, der Anzahl der Phoneme in einem Wort und der Wortlänge in *Frames*, zu einem Klassifizierer zusammengefasst. Die Acoustic Stability und die Hypothesis Density werden im Folgenden etwas genauer vorgestellt.

2.2.1 Acoustic Stability

Die Acoustic Stability – auch A-Stability genannt – wurde erstmals 1996 in [4] und [9] vorgestellt. Für dieses Maß werden alternative Satzhypothesen zu der zu bewertenden Hypothese w_1^N ermittelt. Diese – typischerweise um die 100 – alternativen Hypothesen werden mit verschiedenen Gewichtungen des Sprach- bzw. akustischen Modells berechnet, wodurch unterschiedliche Ergebnisse erzielt werden. Als Konfidenzmaß für jede Worthypothese w_n, $n \in \{1, 2, \ldots, N\}$ aus w_1^N dient nun die Anzahl ihres Vorkommens in den alternativen Hypothesen normalisiert mit der Anzahl der Hypothesen.

$$C_A(w_n) \quad = \quad \frac{1}{M} \cdot \sum_{\{v_1^{N'}\}} \delta(w_n, \mathcal{L}_n(w_1^N, v_1^{N'})). \tag{5}$$

Dabei ist δ die sogenannte Kronecker-Funktion, die den Wert 1 annimmt, wenn beide Argumente übereinstimmen, und sonst 0. $\{v_1^{N'}\}$ ist die Menge der ermittelten Alternativhypothesen mit $|\{v_1^{N'}\}| = M$ und N' sei die Länge der aktuell betrachteten Hypothese. Die Anpassung der beiden jeweils in einem Schritt verglichenen Satzhypothesen wird durch den in Abschnitt 2.1 vorgestellten Levensthein Algorithmus $\mathcal{L}$ vorgenommen.

Wenn eine Worthypothese w_n in vielen oder sogar in allen Alternativen Satzhypothesen vorkommt, bedeutet dies, dass w_n stabil bezüglich unterschiedlicher Skalierung ist und somit wahrscheinlich korrekt erkannt wurde. Ein Wert für $C_A(w_n)$ unwesentlich kleiner als 1 steht somit für eine hohe Sicherheit bei der Erkennung von w_n. Ist der Wert allerdings nahe 0 so besteht eine große Unsicherheit bezüglich dieser Worthypothese.

[5] Der Viterbi-Algorithmus stellt ein effizientes Verfahren zur Bestimmung des besten Pfades durch ein *Hidden Markov Modell* (HMM) dar (s. [10, 13]).

Es ist leicht zu erkennen, dass die an ein Konfidenzmaß gestellten Voraussetzungen von dieser Methode erfüllt werden. Wie gut sich das Verfahren allerdings in der praktischen Anwendung behaupten kann wird in Abschnitt 4 ausführlicher beschrieben.

2.2.2 Hypothesis Density

Die Berechnung der Hypothesis Density beruht direkt auf einem vom Erkennungssystem ermittelten Wortgraphen und kann nicht für die Verwendung mit N-Besten Listen verändert werden. Die Idee beruht darauf, dass ein Spracherkennungssystem bei der Berechnung der besten Satzhypothese unwahrscheinliche Hypothesen verwirft, um den Aufwand in annehmbaren Grenzen zu halten. Das Verwerfen der alternativen Hypothesen geschieht über sogenanntes *Thresholding*, das heißt es wird eine *Grenze* (eng.: threshold) festgelegt und alle Hypothesen, die schlechter bewertet wurden als die Referenzhypothese abzüglich des Thresholds, werden verworfen. Ist in einem Zeitintervall $[\tau_i, t_i]$ die Wahrscheinlichkeit für eine bestimmte Worthypothese $[w_i; \tau_i, t_i]$ sehr hoch so werden nahezu alle anderen Hypothesen verworfen. Für den daraus entstehenden Wortgraphen bedeutet dies, dass es innerhalb dieses Zeitintervalls nur wenige parallele Kanten gibt. Gibt es jedoch viele Worthypothesen mit ähnlichen Wahrscheinlichkeiten, so können nur sehr wenige verworfen werden und es existieren somit im Wortgraphen viele parallele Kanten. Formell ausgedrückt berechnet sich die Hypothesis Density D zu einem Zeitpunkt t' folgendermaßen

$$D(t') \quad = \quad |\{[w;\tau,t] : [w;\tau,t] \in WG \wedge \tau \leq t' \leq t\}|. \tag{6}$$

Wobei WG die Menge aller in dem ermittelten Wortgraphen enthaltenen Worthypothesen ist.

Eine Variation der Hypothesis Density wird in [15] eingeführt. Hier wird nicht die gesamte Anzahl der Hypothesen zu einem Zeitpunkt berechnet, sondern nur die Anzahl der Hypothesen mit unterschiedlichen Wörtern. Dies macht einen Unterschied, da ein Wortgraph in einem Zeitintervall $[\tau_i, t_i]$ das gleiche Wort mit geringfügig unterschiedlichen Start- und Endzeitpunkten mehrmals enthalten kann (siehe Abbildung 4). Dadurch ändert sich Gleichung (6) minimal

$$D'(t') \quad = \quad |\{w : [w;\tau,t] \in WG \wedge \tau \leq t' \leq t\}|. \tag{7}$$

Für das ursprüngliche Verfahren aus [6] wurde die Hypothesendichte zum Startzeitpunkt, zum Endzeitpunkt und die Durchschnittliche Hypothesendichte berechnet. Zudem berechnen die Autoren noch die Hypothesendichte im letzten Frame der Vorgängerhypothese und im ersten Frame der Nachfolgerhypothese und kombinierten diese Merkmale mit der A-Stability und einer Variante der *Word Posterior Wahrscheinlichkeiten* (s. Abschnitt 3). Die durchschnittliche Hypothesis Density für eine Worthypothese berechnet sich mit der Definition der Dichte nach (7) wie folgt

$$C_{D_{avg}}([w;\tau,t]) \quad = \quad \frac{1}{t-\tau+1} \cdot \sum_{t'=\tau}^{t} D'(t'). \tag{8}$$

Für dieses Verfahren gilt, dass ein hoher Wert von $C_{D_{avg}}([w;\tau,t])$ eine hohe Unsicherheit bezüglich der Worthypothese $[w;\tau,t]$ ausdrückt. Ein niedriger Wert hingegen drückt eine relative Sicherheit in der Erkennung aus.

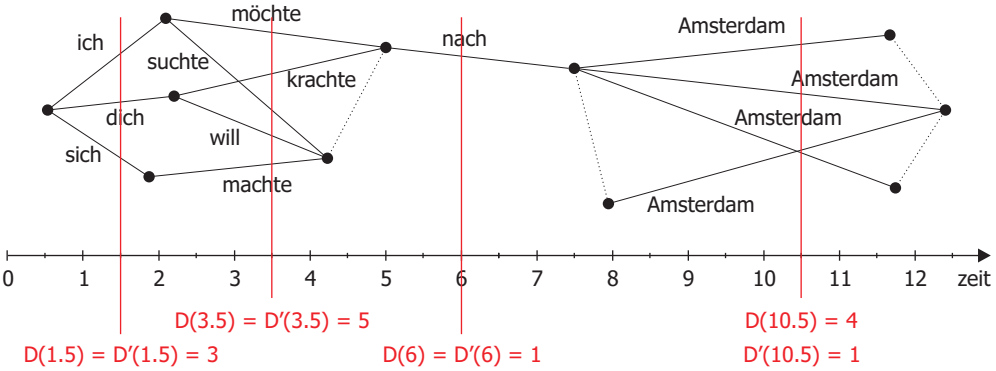


Abbildung 4: Hypothesendichte zu verschiedenen Zeitpunkten in einem einfachen Wortgraphen für den Satz „Ich möchte nach Amsterdam“. Die letzten Worthypothesen für das Wort „Amsterdam“ verdeutlichen den Unterschied zwischen den beiden Dichtemaßen. Ist unter Verwendung von $D(t')$ die Dichte mit 4 sehr hoch, so ist sie, wenn $D'(t')$ verwendet wird, mit 1 extrem niedrig.

2.3 Statistisch motivierte Konfidenzmaße

Unter statistisch motivierten Konfidenzmaßen werden sämtliche Verfahren zusammengefasst, die im weitesten Sinne auf der Berechnung von Wahrscheinlichkeiten beruhen. Durch diese recht grobe Gruppierung ergeben sich abermals mehrere Kategorien. So kann zwischen statistischen Konfidenzmaßen unterschieden werden, die auf Wahrscheinlichkeiten für die Phoneme auf einem Viterbi-Pfad beruhen und solchen, die versuchen die Wahrscheinlichkeiten für ein Wort gegeben die akustische Beobachtung mit Hilfe der Satzwahrscheinlichkeiten zu berechnen.

2.3.1 Phonembasiert

Phonembasierte Ansätze sind vor allem im Bereich der Detektion von *Out Of Vocabulary* (OOV) Worten und damit in der *Äußerungszurückweisung* (eng.: *Utterance Rejection*) zu finden.

In [11] werden 2 Maße vorgestellt, die sich wie folgt berechnen

$$C_{P_{fe}}([ph;\tau,t]_1^N) = \frac{1}{T}\cdot\sum_{i=1}^{N}\sum_{\tau_i\le t'\le t_i}\log[p(ph_i|x_{t'})]. \tag{9}$$

$$C_{P_{pe}}([ph;\tau,t]_1^N) = \frac{1}{N}\cdot\sum_{i=1}^{N}\frac{1}{t_i-\tau_i+1}\cdot\sum_{\tau_i\le t'\le t_i}\log[p(ph_i|x_{t'})]. \tag{10}$$

Dabei ist $[ph;\tau,t]_1^N = [ph_1;\tau_1,t_1],[ph_2;\tau_2,t_2],\ldots,[ph_N;\tau_N,t_N]$ die Folge von Phonemen, die auf dem ermittelten Viterbi-Pfad liegt, mit den zugehörigen Start- und Endzeiten. Die zu der Äußerung gehörende Folge akustischer Beobachtungen ist wie bereits oben x_1^T. Die Wahrscheinlichkeit $p(ph_i|x_{t'})$ wird über die Bayes'sche Regel berechnet[6].

[6] Bayes'sche Regel: $p(g_i|h) = \dfrac{p(h|g_i)\cdot p(g_i)}{\sum_j p(h|g_j)\cdot p(g_j)}$

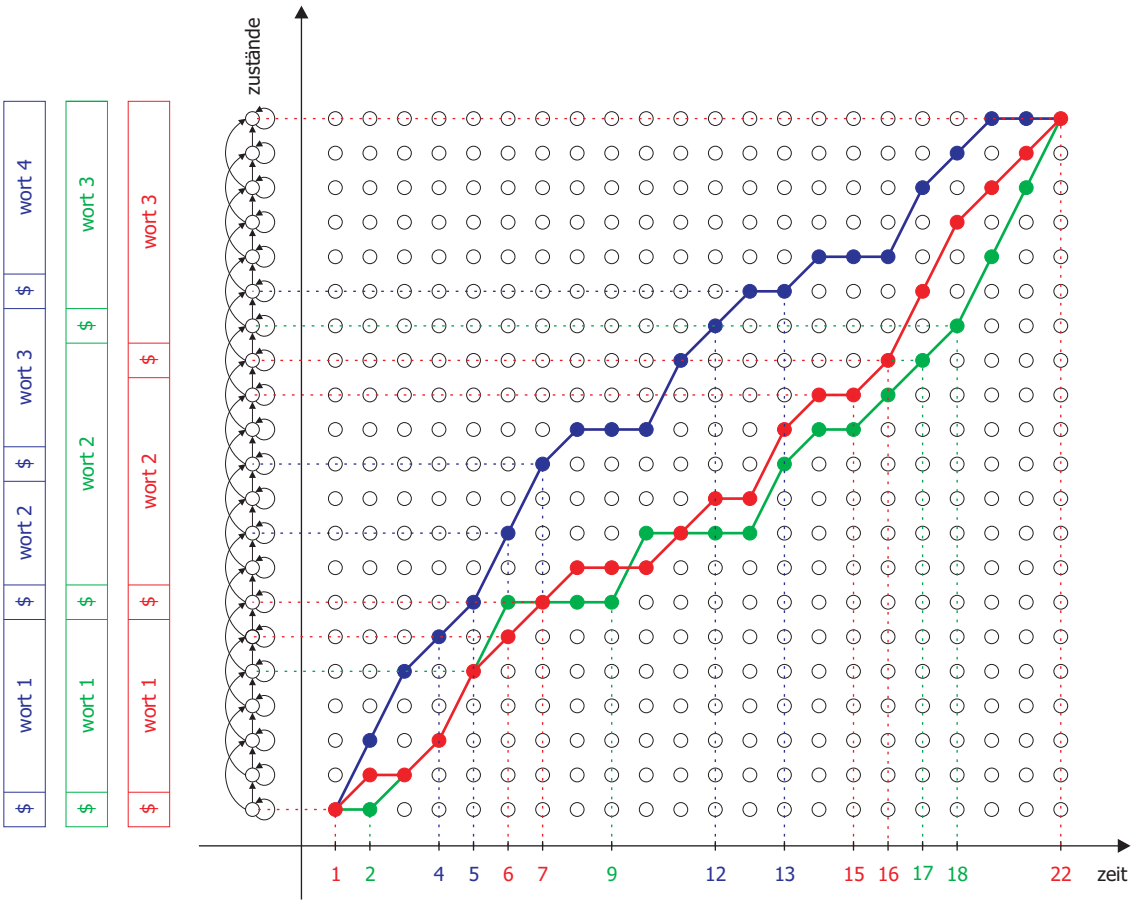


Abbildung 5: Hidden Markov Model (HMM) mit 3 alternativen Pfaden (Hypothesen) durch das Modell. Die favorisierte (rote) Hypothese wird vom Viterbi-Algorithmus ermittelt. Die gestrichelten Linien repräsentieren die Start- und Endzeiten der Wörter. $ steht für eine Pause [14].

Die beiden Konfidenzmaße unterscheiden sich nur in der Normalisierung der Funktionen. So wird in $C_{P_{fe}}$ mit der Anzahl der akustischen Beobachtungen (Frames) normalisiert und bei $C_{P_{pe}}$ mit der Anzahl der Phoneme und der Länge des jeweiligen Phonems. Nach der Definition dieser Verfahren aus [11] berechnen sie nur die Konfidenz für eine gesamte Äußerung, können jedoch leicht so angepasst werden, dass sie auch als Konfidenzmaß für einzelne Wörter verwendet werden können.

Ein anderes Verfahren aus [1] vergleicht die Wahrscheinlichkeit für eine akustische Beobachtung, gegeben das erkannte Wort mit der Wahrscheinlichkeit für die akustische Beobachtung, gegeben die Folge von Phonemen ph^*, die am besten auf die Beobachtung passt, ohne dass diese ein sinnvolles Wort ergeben muss.

$$C_{P_{ph}}([w;\tau,t]) \quad = \quad \frac{p(x_\tau^t|w)}{p(x_\tau^t|ph^*)}. \tag{11}$$

Die Vermutung, die hinter diesem Ansatz steckt, ist die, dass, wenn die Wahrscheinlichkeit für ein sinnloses OOV Wort höher ist als die für das erkannte Wort, es wahrscheinlich ist, dass dieses Wort falsch erkannt wurde.

Mit Hilfe dieses Verfahren kann auch eine Konfidenz für eine Worthypothese eines Wortgraphen ermittelt werden, die zu einem bestimmten Zeitpunkt die einzige Worthypothese im Wortgraphen ist. Ein Beispiel für eine solche Situation ist die Hypothese $[nach; 5, 7.5]$ in Abbildung 4. Verfahren, die auf Wörtern beruhen ordnen dieser Worthypothese automatisch best mögliche Konfidenz zu, da sie im entsprechenden Zeitraum die einzige Hypothese ist. Kam in der zu erkennenden Äußerung allerdings ein OOV Wort vor und wurde durch ein Wort aus dem Vokabular substituiert, so kann dies nur mittels eines Konfidenzmaßes auf Phonembasis, wie zum Beispiel $C_{P_{ph}}([w; \tau, t])$, repräsentiert werden.

2.3.2 Word Posterior Wahrscheinlichkeiten

Die Verwendung der bedingten Wahrscheinlichkeit für eine Worthypothese gegeben die akustische Beobachtung als Konfidenzmaß für diese Hypothese ist an und für sich ein recht naheliegendes Verfahren, da diese Wahrscheinlichkeit genau das, was ein Konfidenzmaß ausdrücken soll, darstellt. Die Wahrscheinlichkeit $p([w; \tau, t]|x_1^T)$ wird im folgenden als Word Posterior Wahrscheinlichkeit bezeichnet. Die Schwierigkeit, die sich bei der Berechnung dieser Wahrscheinlichkeit ergibt liegt zum einen darin, das in Gleichung (1) nicht mit $p(x_1^T)$ normalisiert wird. Aber selbst wenn diese Wahrscheinlichkeit gegeben wäre, wäre die Berechnung der Word Posterior Wahrscheinlichkeiten für eine Worthypothese über die Summe aller Satzhypothesen, welche diese Worthypothese enthalten zu komplex, um direkt ermittelt werden zu können.

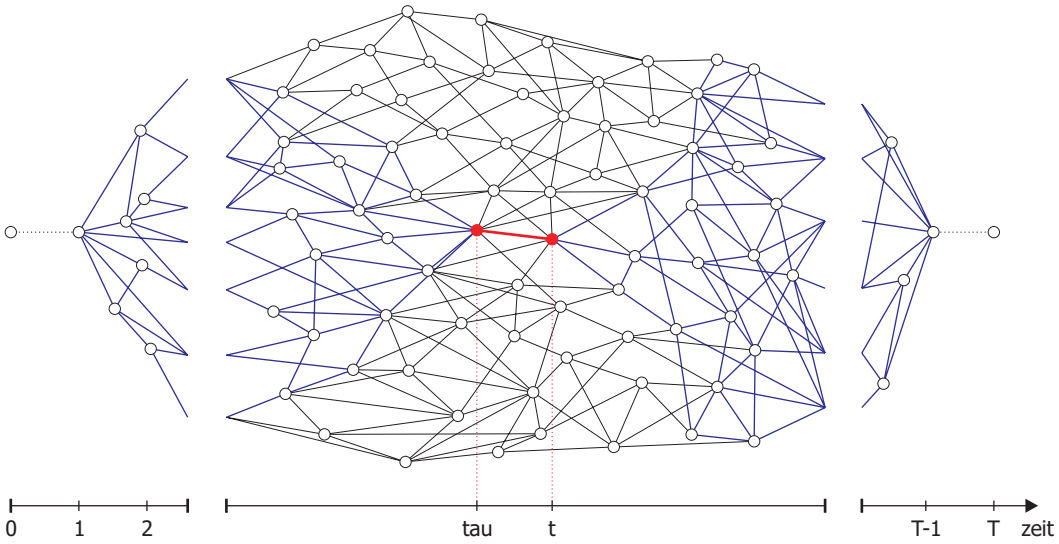


Abbildung 6: Wortgraph mit einer zu bewertenden Worthypothese $[w; \tau, t]$ (rot). Um die Posterior Wahrscheinlichkeit $p([w; \tau, t]|x_1^T)$ zu berechnen, müssen die Wahrscheinlichkeiten aller Satzhypothesen (blau), welche die betrachtete Hypothese enthalten, aufsummiert werden. Auf die Unterscheidung zwischen „normalen“ Kanten und Pausenkanten wurde in diesem Beispiel verzichtet.

$$
\begin{aligned}
p([w;\tau,t]|x_1^T) &= \sum_N \sum_{\substack{[w;\tau,t]_1^N: \\ \exists n\in\{1,2,\ldots,N\}: \\ [w_n;\tau_n,t_n]=[w;\tau,t]}} p([w;\tau,t]_1^N|x_1^T) \\
&= \sum_N \sum_{\substack{[w;\tau,t]_1^N: \\ \exists n\in\{1,2,\ldots,N\}: \\ [w_n;\tau_n,t_n]=[w;\tau,t]}} \frac{p(x_1^T|[w;\tau,t]_1^N)\cdot p(w_1^N)}{p(x_1^T)} \\
&= \sum_N \sum_{\substack{[w;\tau,t]_1^N: \\ \exists n\in\{1,2,\ldots,N\}: \\ [w_n;\tau_n,t_n]=[w;\tau,t]}} \frac{p(x_1^T,[w;\tau,t]_1^N)}{p(x_1^T)} \\
&= \sum_N \sum_{\substack{[w;\tau,t]_1^N: \\ \exists n\in\{1,2,\ldots,N\}: \\ [w_n;\tau_n,t_n]=[w;\tau,t]}} \frac{\prod_{n=1}^N p(x_{\tau_n}^{t_n}|w_n)\cdot p(w_n|w_1^{n-1})}{p(x_1^T)}. \qquad (12)
\end{aligned}
$$

Die beiden zentralen Probleme bei der Verwendung der Word Posterior Wahrscheinlichkeiten als Konfidenzmaß liegen nun, wie bereits erwähnt in der Abschätzung von $p(x_1^T)$ und vor allem in der effizienten Berechnung von Gleichung (12).

Abschätzung der akustischen Wahrscheinlichkeit Ein Ansatz, der insbesondere von Willet in [18, 19] verfolgt wird ist, die Abschätzung von $p(x_1^T)$ mit Hilfe des zugrunde liegenden HMM's

$$
p(x_i) \approx \sum_{i'=1}^{S} p(x_i|s_{i'})\cdot p(s_{i'}). \qquad (13)
$$

$$
p(x_\tau^t) \approx \left(\prod_{i=\tau}^{t} p(x_i)\right)\cdot t_{avg}^{t-\tau}. \qquad (14)
$$

Die Summe in Gleichung (13) geht über sämtliche Zustände $s_1^S = s_1, s_2, \ldots, s_S$ des HMM's. Nach [19] kann die a-priori Wahrscheinlichkeit sich in einem Zustand des HMM's zu befinden $p(s_{i'})$ einfach auf Basis der Trainingsdaten approximiert werden. Die gesamte Gleichung (13) kann, abhängig vom jeweils verwendeten Erkennungssystem, vereinfacht und effizient berechnet werden [18]. Die a-priori Wahrscheinlichkeit für eine akustische Beobachtung wird (s. Gleichung (14)) durch das Produkt über die Einzelwahrscheinlichkeiten unter Annahme der stochastischen Unabhängigkeit approximiert und mit t_{avg} – der durchschnittlichen Übergangswahrscheinlichkeit des HMM's – normalisiert.

Mit Hilfe dieser Abschätzungen wird in [19] folgendes Konfidenzmaß definiert

$$
\begin{aligned}
C_{W_{est}}([w;\tau,t]) &= p(w|x_\tau^t) \\
&= \frac{p(x_\tau^t|w)\cdot p(w)}{p(x_\tau^t)}. \qquad (15)
\end{aligned}
$$

Für die Wahrscheinlichkeit $p(x_\tau^t|w)$ wird der vom Erkennungssystem approximierte Wert verwendet und $p(w)$ wird über das Sprachmodell ermittelt.

Die Kombination dieses Verfahrens mit Konfidenzmaßen auf Basis des Sprachmodells liefert nach [19] schlechtere Ergebnisse als die bereits in Abschnitt 2.2.1 und 2.2.2 vorgestellten heuristischen Verfahren, lässt sich jedoch relativ effizient berechnen und ist damit auch für den Einsatz in Echtzeitapplikationen geeignet.

Forward-Backward-Algorithmus Wie in Abbildung 6 zu sehen, kann durch die Einschränkung, dass die Worthypothese in den Satzhypothesen enthalten sein muss, nur ein kleiner Teil aller Worthypothesen ausgeschlossen werden. Ein anderer Ansatz, der in [6] mit *Link Probability* bezeichnet wird, ist die effiziente Berechnung der Word Posterior Wahrscheinlichkeit auf Wortgraphen mit dem bekannten Forward-Backward-Algorithmus [10]. Dieser Ansatz wird auch in [14, 16, 17] verfolgt und wird in Abschnitt 3 genauer beschrieben.

3 Berechnung von Word Posterior Wahrscheinlichkeiten

3.1 Wortgraphen basierte Definition

Die Idee in [6] und [17] ist, die Anwendung des bekannten *Forward-Backward-Algorithmus* zur Berechnung der Word Posterior Wahrscheinlichkeiten. Hierzu wird das Problem der Berechnung der Summe aus Gleichung (12) in 2 Teilprobleme zerlegt. Zum einen in die Wahrscheinlichkeit bis zu dem betrachteten Wort zu kommen und zum anderen in die Wahrscheinlichkeit vom Ende der betrachteten Worthypothese bis zum Satzende zu kommen. Diese beiden Wahrscheinlichkeiten sind die *Forward-* und die *Backward-Wahrscheinlichkeit*. Diese beiden Wahrscheinlichkeiten lassen sich, wie weiter unten beschrieben rekursiv darstellen und somit recht effizient mittels dynamischer Programmierung lösen.

Forward- und Backward-Wahrscheinlichkeiten Zur Berechnung der Word Posterior Wahrscheinlichkeit auf Wortgraphen mittels des Forward-Backward-Algorithmus muss man sich die Forward- und eine Backward-Wahrscheinlichkeit definieren. Diese Wahrscheinlichkeiten hängen von dem verwendeten m-gram Sprachmodell ab, da bei beiden Wahrscheinlichkeiten die Vorgänger bzw. die Nachfolger des betrachteten Wortes beachtet werden müssen. Die Forward-Wahrscheinlichkeit $\Phi([w;\tau,t], h_w^{(m)})$ drückt aus, dass die letzte Worthypothese einer Hypothesenfolge der Länge n $[w;\tau,t]$ ist und bezieht ebenfalls die *Historie* der Hypothese mit ein. Wobei n die Position der Worthypothese in der besten Satzhypothese ist. Betrachten wir erst einmal den einfachen Fall eines *uni*-grams (d.h. es wird keine Historie betrachtet). Dann definiert sich die Forward-Wahrscheinlichkeit $\Phi_{uni}([w;\tau,t])$ durch

$$\begin{aligned}\Phi_{uni}([w;\tau,t]) &= \sum_n \sum_{\substack{[w;\tau,t]_1^n: \\ [w_n;\tau_n,t_n]=[w;\tau,t]}} p(x_1^{t_n}, [w;\tau,t]_1^n) \\ &= \sum_n \sum_{\substack{[w;\tau,t]_1^n: \\ [w_n;\tau_n,t_n]=[w;\tau,t]}} \prod_{j=1}^{n} p(x_{\tau_j}^{t_j}|w_j)\cdot p(w_j). \end{aligned} \tag{16}$$

Die Backward-Wahrscheinlichkeit $\Psi([w;\tau,t], f_w^{(m)})$ stellt die Wahrscheinlichkeit dar, dass $[w;\tau,t]$ die erste Worthypothese in einer Hypothesenfolge der Länge n ist. Wobei $n = N-i+1$ gilt, wenn $[w;\tau,t]$ die $i$te Worthypothese in einer Satzhypothese der Länge N ist.

Im Gegensatz zur Forward-Wahrscheinlichkeit muss bei der Backward-Wahrscheinlichkeit nicht die Historie eines Wortes mit einbezogen werden, sondern seine Nachfolger – auch die *Zukunft* oder *Future* eines Wortes genannt. $f_1^{m-1} = w_{i+1}, w_{i+2}, \ldots, w_{\min\{N,i+m-1\}}$ sind somit, analog zur Definition der Historie aus 2, die $(m-1)$ Nachfolger von w_i. Die Backward-Wahrscheinlichkeit für ein *uni*-gram definiert sich somit ganz ähnlich wie die Forward-Wahrscheinlichkeit in Gleichung (16)

$$\begin{aligned}\Psi_{uni}([w;\tau,t]) &= \sum_n \sum_{\substack{[w;\tau,t]_1^n: \\ [w_1;\tau_1,t_1]=[w;\tau,t]}} p(x_\tau^{t_n}, [w;\tau,t]_1^n) \\ &= \sum_n \sum_{\substack{[w;\tau,t]_1^n: \\ [w_1;\tau_1,t_1]=[w;\tau,t]}} \prod_{j=1}^{n} p(x_{\tau_j}^{t_j}|w_j) \cdot p(w_j). \end{aligned} \tag{17}$$

Es ist zu beachten, dass für die Backward-Wahrscheinlichkeit einer Hypothese $[w;\tau,t]$ auch die akustischen Beobachtungen, angefangen beim Beginn der Hypothese τ bis zur letzten Beobachtung T, einbezogen werden müssen. Somit gilt in Gleichung (17) $t_n = T$.

Durch herausnehmen des letzten Summanden aus der Summe in Gleichung (16) kann die Forward-Wahrscheinlichkeit in eine rekursive Definition umgeschrieben werden

$$\begin{aligned}\Phi_{uni}([w;\tau,t]) &= p(x_\tau^t|w) \cdot p(w) \cdot \sum_n \sum_{\substack{[w;\tau,t]_1^{n-1}: \\ t_{n-1}=\tau-1}} \prod_{j=1}^{n-1} p(x_{\tau_j}^{t_j}|w_j) \cdot p(w_j) \\ &= p(x_\tau^t|w) \cdot p(w) \cdot \sum_{w'} \sum_{\tau'} \Phi_{uni}([w';\tau',\tau-1]). \end{aligned} \tag{18}$$

Eine ähnliche Umformung lässt sich auch für Gleichung (17) durchführen, wodurch sich

$$\Psi_{uni}([w;\tau,t]) = p(x_\tau^t|w) \cdot p(w) \cdot \sum_{w'} \sum_{t'} \Psi_{uni}([w';t+1,t']). \tag{19}$$

ergibt.

Bei der allgemeineren Betrachtung eines m-gram Sprachmodells verändern sich die beiden Definitionen (16) und (17) durch die Beachtung der Historie bzw. der Zukunft

geringfügig

$$
\begin{aligned}
\Phi([w;\tau,t],h_2^{m-1}) &= \sum_n \sum_{\substack{[w;\tau,t]_1^n: \\ [w_n;\tau_n,t_n]=[w;\tau,t], \\ w_{n-m+2}^{n-1}=h_2^{m-1}}} p(x_1^{t_n},[w;\tau,t]_1^n) \\
&= \sum_n \sum_{\substack{[w;\tau,t]_1^n: \\ [w_n;\tau_n,t_n]=[w;\tau,t], \\ w_{n-m+2}^{n-1}=h_2^{m-1}}} \prod_{j=1}^{n} p(x_{\tau_j}^{t_j}|w_j)\cdot p(w_j|w_{j-m+1}^{j-1}) \\
&= p(x_\tau^t|w)\cdot \sum_n \sum_{\substack{[w;\tau,t]_1^{n-1}: \\ t_{n-1}=\tau-1, \\ w_{n-m+2}^{n-1}=h_2^{m-1}}} p(w|h_1^{m-1}) \\
&\quad \cdot \prod_{j=1}^{n-1} p(x_{\tau_j}^{t_j}|w_j)\cdot p(w_j|w_{j-m+1}^{j-1}) \\
&= p(x_\tau^t|w)\cdot \sum_{h_1} p(w|h_1^{m-1}) \\
&\quad \cdot \sum_{\tau'} \Phi([h_{m-1};\tau',\tau-1],h_1^{m-2}). \qquad (20)
\end{aligned}
$$

Die Definition der Backward-Wahrscheinlichkeit mit Betrachtung der Zukunft eines Wortes ist ungleich komplizierter, da das verwendete m-gram Sprachmodell vorwärts gerichtet ist und somit nur die Historie eines Wortes bereitstellt. So muss die Zukunft eines Wortes w_i rückwirkend mit einbezogen werden. Dies geschieht, indem statt der eigentlichen Zukunft $\prod_{j=1}^{n} p(w_j|f_1^{m-2}) = \prod_{j=1}^{n} p(w_j|w_{j+1}^{\min\{N-1,i+m-2\}})$ immer die Historie ab dem $m$ten Wort $\prod_{k=m}^{n} p(w_k|w_{\max\{1,k-m+1\}}^{k-2}) = \prod_{k=m}^{n} p(w_k|h_2^{m-1})$ verwendet wird. Durch diesen Trick kann die eine leicht abgewandelte Backward-Wahrscheinlichkeit $\tilde{\Psi}([w;\tau,t],f_1^{m-2})$ wie folgt definiert werden, die allerdings noch kein Sprachmodell enthält

$$
\begin{aligned}
\tilde{\Psi}([w;\tau,t],f_1^{m-2}) &= \sum_n \sum_{\substack{[w;\tau,t]_1^n: \\ [w_1;\tau_1,t_1]=[w;\tau,t], \\ w_2^{m-1}=f_1^{m-2}}} p(x_\tau^{t_n},[w;\tau,t]_1^n) \\
&= \sum_n \sum_{\substack{[w;\tau,t]_1^n: \\ [w_1;\tau_1,t_1]=[w;\tau,t], \\ w_2^{m-1}=f_1^{m-2}}} \prod_{j=1}^{n} p(x_{\tau_j}^{t_j}|w_j)\cdot p(w_j|f_1^{m-2}) \\
&= \sum_n \sum_{\substack{[w;\tau,t]_1^n: \\ [w_1;\tau_1,t_1]=[w;\tau,t], \\ w_2^{m-1}=f_1^{m-2}}} \prod_{j=1}^{n} p(x_{\tau_j}^{t_j}|w_j)\cdot \prod_{k=m}^{n} p(w_k|w_{k-m+1}^{j-1}). \qquad (21)
\end{aligned}
$$

Auch Gleichung (21) kann rekursiv berechnet werden, wobei die Backward-Wahrschein-

lichkeit in der folgenden Definition in absteigender Reihenfolge berechnet wird

$$\begin{aligned}\tilde{\Psi}([w;\tau,t],f_1^{m-2}) &= p(x_\tau^t|w)\cdot\sum_n \sum_{\substack{[w;\tau,t]_1^{n-1}:\\ \tau_1=t+1,\\ w_2^m=f_1^{m-1}}} p(w|f_1^{m-2}) \\ &\quad\cdot\prod_{j=1}^{n-1} p(x_{\tau_j}^{t_j}|w_j)\cdot p(w_j|f_2^{m-1}) \\ &= p(x_\tau^t|w)\cdot\sum_{f_{m-1}} p(f_{m-1}|w,f_1^{m-2}) \\ &\quad\cdot\sum_{t'}\Phi([f_1;t+1,t'],f_2^{m-1}). \end{aligned} \tag{22}$$

Kombination der Wahrscheinlichkeiten Im einfachen Fall eines *uni*gram Sprachmodells kann anhand der Gleichungen (19) und (18) die Word Posterior Wahrscheinlichkeit als Produkt der Forward- und Backward-Wahrscheinlichkeiten berechnet werden

$$p_{uni}([w;\tau,t]|x_1^T) = \frac{\Phi_{uni}([w;\tau,t])\cdot\Psi_{uni}([w;\tau,t])}{p(x_1^T)\cdot p(w)\cdot p(x_\tau^t|w)}. \tag{23}$$

Bei der Word Posterior Wahrscheinlichkeit für ein m-gram Sprachmodell müssen die in Gleichung (22) fehlenden Sprachmodell Wahrscheinlichkeiten noch nachträglich ergänzt werden. Somit ergibt sich mit den Definitionen (20) und (22) die Wahrscheinlichkeit für eine Worthypothese in [17] wie folgt

$$\begin{aligned} p([w;\tau,t]|x_1^T) &= \sum_{h_2^{m-1}}\sum_{f_1^{m-2}} \frac{\Phi([w;\tau,t],h_2^{m-1})\cdot\Psi([w;\tau,t],f_1^{m-2})}{p(x_1^T)\cdot p(x_\tau^t|w)} \\ &\quad\cdot\prod_{n=1}^{m-2} p(f_n|h_{n+1}^{m-1},w,f_1^{n-1}). \end{aligned} \tag{24}$$

Das Produkt in Gleichung (24) stellt die in (21) und (22) noch fehlenden Sprachmodell Wahrscheinlichkeiten dar.

In den Gleichungen (23) und (24) wird durch die akustische Wahrscheinlichkeit geteilt, da diese sowohl durch (18) als auch durch (19) in die Gleichungen mit einfließt. Die Wahrscheinlichkeit für die akustische Beobachtung $p(x_1^T)$ kann ebenfalls mit Hilfe der Forward- und Backward-Wahrscheinlichkeiten gebildet werden

$$\begin{aligned} p(x_1^T) &= \sum_{i=1}^{N}\sum_{h_2^{m-1}}\sum_\tau \Phi([w;\tau,T],h_2^{m-1}) \\ &= \sum_{i=1}^{N}\sum_{f_1^{m-2}}\sum_t \Psi([w;1,t],f_1^{m-2}) \\ &\quad\cdot\left[\prod_{n=1}^{m-2} p(f_n|w,f_1^{n-1})\right]\cdot p(w). \end{aligned} \tag{25}$$

Die somit berechnete Word Posterior Wahrscheinlichkeit kann direkt als Konfidenzmaß genutzt werden.

$$C_W([w;\tau,t]) = p([w;\tau,t]|x_1^T). \tag{26}$$

Optimierung der Word Posterior Wahrscheinlichkeiten Wie sich in den Experimenten (s. Abschnitt 4.2) gezeigt hat, liefert die reine Word Posterior Wahrscheinlichkeit nach Gleichung (24) keine sonderlich gute Leistung als Konfidenzmaß. Um das Verfahren weiter zu verbessern werden in [17] noch 2 Veränderungen in den Berechnungen durchgeführt. Die Probleme des bisherigen Verfahrens liegen zum einen in den ungewichteten Wahrscheinlichkeiten des Sprach- und des akustischen Modells. Zum anderen werden viele Worthypothesen, die sich nur minimal in Anfangs- und Endzeit unterscheiden, durch die strikte Vorgabe dieser Werte nicht mit in die Berechnung einbezogen, obwohl diese im Prinzip die gleiche Hypothese darstellen. Dadurch teilt sich die Word Posterior Wahrscheinlichkeit unter allen ähnlichen Hypothesen auf, was wiederum zu einer schlechteren Klassifizierung führt.

Die Anpassung der Wahrscheinlichkeiten geschieht ähnlich wie, während des Erkennungsprozesses über *Skalierungsfaktoren* (eng.: *Scaling Factors*). Es wird in [17] sowohl ein Skalierungsfaktor für die Wahrscheinlichkeiten des Sprachmodells verwendet, dieser wird mit β bezeichnet, als auch einer für das akustische Modell, welcher α genannt wird. Diese beiden Faktoren haben einen großen Einfluss auf die Güte der Word Posterior Wahrscheinlichkeiten und werden anhand von Testmustern für die eigentliche Bewertung optimiert. Gleichung (20) mit den Skalierungsfaktoren sieht dann folgendermaßen aus

$$\begin{aligned}\Phi([w;\tau,t],h_2^{m-1}) &= p(x_\tau^t|w)^\alpha \cdot \sum_{h_1} p(w|h_1^{m-1})^\beta \\ &\quad \cdot \sum_{\tau'} \Phi([h_{m-1};\tau',\tau-1],h_1^{m-2}). \end{aligned} \tag{27}$$

Den Einfluss der Faktoren auf die Güte der Word Posterior Wahrscheinlichkeiten als Konfidenzmaß zeigt Abbildung 7.

Das zweite Problem der Hypothesen, die das gleiche Wort darstellen und parallel zueinander im Wortgraphen verlaufen, lässt sich auf verschiedene Weisen lösen. In [17] werden dazu 3 Ansätze vorgestellt. So können zum Beispiel alle Hypothesen in den Konfidenzwert für eine Hypothese mit einbezogen werden, die das gleiche Wort darstellen und mit der betrachteten Hypothese überlappen

$$C_{W_{sec}}([w;\tau,t]) = \sum_{\substack{[w;\tau',t']:\\ \{\tau,\ldots,t\}\cap\{\tau',\ldots,t'\}\neq\emptyset}} p([w;\tau',t']|x_1^T). \tag{28}$$

Der Nachteil dieses Verfahrens ist, dass die Summe der verschiedenen Word Posterior Wahrscheinlichkeiten, die in $C_{W_{sec}}$ aufsummiert werden, größer als 1 werden kann und somit keine Normierung und auch keine Vergleichbarkeit des Konfidenzmaßes mehr vorhanden wäre. Dies ist der Fall, da durch das Betrachten aller Hypothesen, die mit der zu bewertenden überlappen, die zusätzlich betrachteten Hypothesen selbst nicht überlappen müssen und daher über disjunkte Zeiträume aufsummiert wird. Als Lösung dieses Problems kann die Summe auf Hypothesen beschränkt werden, deren Zeitrahmen paarweise nicht disjunkt sind. Zu diesem Zweck wird für alle Hypothesen ein Zeitpunkt der Referenzhypothese als gemeinsam festgelegt. Im Falle von $C_{W_{med}}$ ist dies $lceil(\tau+t)/2\rceil$, also die Mitte der Referenzhypothese

$$C_{W_{med}}([w;\tau,t]) = \sum_{\substack{[w;\tau',t']:\\ \tau'\leq\lceil(\tau+t)/2\rceil\leq t'}} p([w;\tau',t']|x_1^T). \tag{29}$$

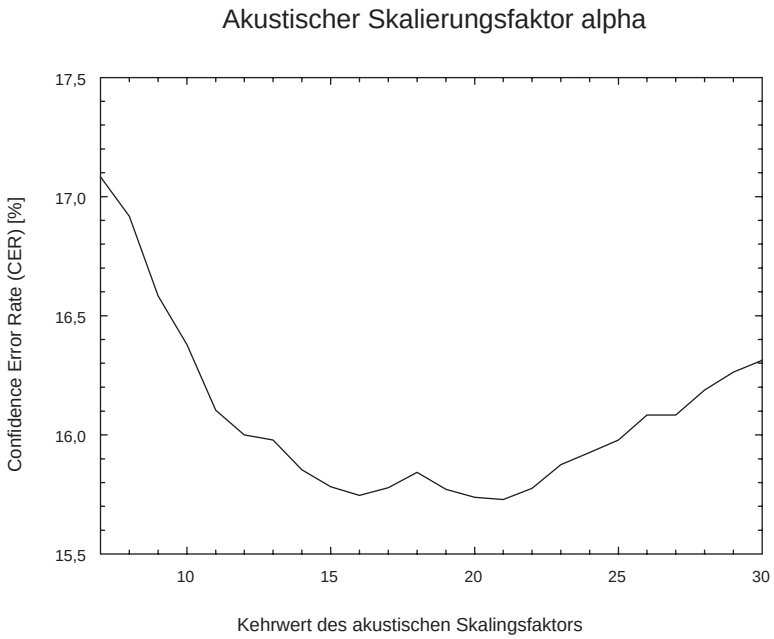


Abbildung 7: *Konfidenz Fehlerraten* (CER) (Die CER ist als der Quotient aus der Anzahl der korrekt markierten Wörter und der Gesamtzahl an Wörtern in einem Satz definiert. S. 4.2) für verschiedene akustische Skalierungsfaktoren [17]. Die Werte wurden auf dem Verbmobil Datensatz mit der einfachen Word Posterior Wahrscheinlichkeit $p([w,\tau,t]|x_1^T)$ als Konfidenzmaß ermittelt.

Als letzte Verbesserung wird in [17] vorgeschlagen, nicht immer nur die Mitte der zu bewertenden Hypothese als gemeinsamen Zeitpunkt festzulegen, sondern die Summe über diesen Zeitpunkt zu maximieren

$$C_{W_{max}}([w;\tau,t]) \quad = \quad \max_{t_{max}\in\{\tau,\ldots,t\}} \sum_{\substack{[w;\tau',t']:\\ \tau'\leq t_{max}\leq t'}} p([w;\tau',t']|x_1^T). \tag{30}$$

Speziell für die Berechnung der Word Posterior Wahrscheinlichkeiten auf Wortgraphen ist es desweiteren wichtig redundante Pausenkanten aus dem Graphen zu entfernen. Dabei wird eine Pausenkante als redundant bezeichnet, wenn sie durch eine Sequenz mehrerer Pausenkanten dargestellt werden kann. Somit enthalten solche Kanten keine zusätzliche Information, beeinflussen allerdings die Wahrscheinlichkeiten sämtlicher Hypothesen, die mit ihr überlappen, da die Summe aller Wahrscheinlichkeiten zu einem Zeitpunkt 1 ergeben muss.

Das Entfernen der redundanten Kanten hat desweiteren den Effekt, dass die Berechnung der Wahrscheinlichkeiten effizienter durchgeführt werden kann, insbesondere, wenn der ursprüngliche Wortgraph eine große Anzahl an Pausenkanten enthielt.

3.2 N-Besten Listen basierte Definition

Desweiteren wird in [17] noch eine Definition der Word Posterior Wahrscheinlichkeiten auf Basis von N-Besten Listen vorgestellt. Bei diesem Verfahren wird die Wahrscheinlichkeit ähnlich wie in Gleichung (12) über die Summe aller Satzhypothesen, welche die betrachtete

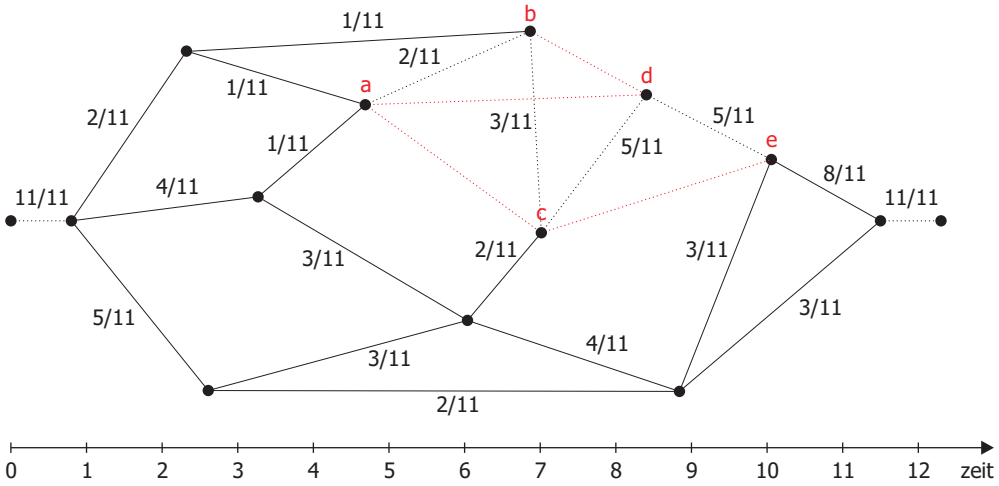


Abbildung 8: Wortgraph aus Abbildung 2 aus dem die redundanten Pausenkanten nach [17] entfernt wurden. An den Kanten stehen die Word Posterior Wahrscheinlichkeiten für die jeweiligen Hypothesen. Die Werte selbst entsprechen der Anzahl der durch die jeweilige Hypothese laufenden Satzhypothesen. Da sich die Anzahl der Satzhypothesen durch Entfernen der Pausenkanten verringert, müssen diese Wahrscheinlichkeiten ebenfalls angepasst werden. In diesem Beispiel wurde die Anzahl der mit dem Wortgraph dargestellten Satzhypothesen von 18 auf 11 reduziert.

Worthypothese enthalten, gebildet. Aufgrund der relativ geringen Anzahl der in der N-Besten Liste gespeicherten Satzhypothesen, ist in diesem Falls diese Summe recht leicht zu berechnen.

$$p(w_m|x_1^T, w_1^M, \mathcal{L}) = \frac{\sum_N \sum_{\{v_1^N\}} p(x_1^T|v_1^N)^\alpha \cdot p(v_1^N)^\beta \cdot \delta(w_m, \mathcal{L}_m(w_1^M, v_1^N))}{\sum_N \sum_{\{v_1^N\}} p(x_1^T|v_1^N)^\alpha \cdot p(v_1^N)^\beta}. \quad (31)$$

δ ist wie bereits oben die Kronecker-Funktion. α und β sind wiederum die Skalierungsfaktoren für das Sprach- und das akustische Modell. Zur Anpassung der verschiedenen Satzhypothesen wird der in Abschnitt 2.1 vorgestellte Levensthein Algorithmus $\mathcal{L}$ verwendet. Die so berechnete Wahrscheinlichkeit kann ohne zusätzliche Veränderungen direkt als Konfidenzmaß genutzt werden und somit definiert sich C_N durch

$$C_N([w;\tau,t]) = p(w|x_1^T, w_1^M, \mathcal{L}). \quad (32)$$

mit

$$[w;\tau,t]_1^M = [w_1;\tau_1,t_1],[w_2;\tau_2,t_2],\ldots,[w_{m-1};\tau_{m-1},t_{m-1}], \\ [w;\tau,t],[w_{m+1};\tau_{m+1},t_{m+1}],\ldots,[w_M;\tau_M,t_M].$$

Wie in Abschnitt 4 zu sehen liefert dieses Verfahren auf den verschiedenen Datensätzen Resultate, die in etwa mit denen des Forward-Backward-Verfahrens vergleichbar sind.

4 Vergleich verschiedener Verfahren

Um die vorgestellten Verfahren mit ihren Vor- und Nachteilen vergleichen zu können, werden die Algorithmen auf unterschiedlichen Test-Datensätzen ausgeführt. Dazu wird ein

Post-Klassifizierer konstruiert, der abhängig von dem Konfidenzwert eines Wortes dieses Wort entweder als *Korrekt* oder *Falsch* markiert. Dieser Post-Klassifizierer muss anhand eines *Kreuzvalidierungs* Datensatzes[7] (eng.: *Cross-Validation Corpus*) optimiert werden.

Manche Evaluierungsmaße werden mit der sogenannten Baseline verglichen. Die Baseline ist als ein Konfidenzmaß definiert, das davon ausgeht, dass der Output des Klassifizierers vollkommen korrekt ist. Also tagt die Baseline alle Wörter der Satzhypothesen einfach als „korrekt“. Die hier vorgestellten Experimente stammen aus [14] und [17].

Die Datensätze selbst werden in Abschnitt 4.1 und Tabelle 1 kurz vorgestellt, um bewerten zu können was für Anforderungen ein bestimmter Korpus an ein Erkennungssystem und ein Konfidenzmaß stellt. Zur eigentlichen Bewertung der Verfahren gibt es verschiedene Kriterien, die alle mehr oder weniger nützlich sind, abhängig von der Art des betrachteten Maßes und der gesuchten Information.

4.1 Datensätze

Im Folgenden werden kurz die Datensätze vorgestellt, auf denen anschließend in Abschnitt 4.3 die verschiedenen Verfahren getestet werden. Die Verfahren wurden in [17] und [14] auf insgesamt 5 Testdatensätze untersucht, die in Tabelle 1 kurz beschrieben sind.

Tabelle 1: Grobe Charakterisierung der 5 Testdatensätze.

Datensatz	Vokabular-größe	Sprache	Dialoge	Aufnahme-bedingungen
ARISE	985	Holländisch	Mensch – Maschine	Telefon
Verbmobil	7128	Deutsch	Mensch – Mensch	Studio
NAB 20k	19987	Englisch	Monolog	Studio
NAB 64k	64736	Englisch	Monolog	Studio
Broadc. News	65491	Englisch	Monolog	Fernsehen / Radio

ARISE Der holländische ARISE Korpus ist mit ungefähr 1.000 Wörtern der kleinste Datensatz im Test. Er besteht aus Anrufen bei einem automatischen Auskunftssystem der holländischen Bahn. Das heißt die Aufnahmen wurden über das Telefon aufgenommen und bestehen aus Mensch – Maschine Dialogen.

Verbmobil Der Verbmobil Datensatz beinhaltet im Studio aufgenommene, spontane Mensch – Mensch Dialoge aus der Terminplanungsdomäne. Mit ca. 7.000 Wörter gehört er ebenfalls zu den kleineren Datensätzen.

NAB Die beiden *North American Business* (NAB) Korpora bestehen aus englischen Zeitungsartikeln, die im Studio gelesen wurden. Der Unterschied zwischen NAB20k und NAB64k besteht darin, das einmal ein Erkenner mit einem Vokabular von ca. 20.000 Wörtern zur Erkennung verwendet wurde und bei NAB64k ein Erkennungssystem mit einem Vokabular von ca. 64.000 Wörtern.

[7] Hierzu wird ein sowohl von den Trainingsdaten als auch von den Testdaten distinkter Teil des Datensatzes genutzt, damit keine Kenntnis über diese Daten in den Klassifizierer einfließt und dadurch die experimentellen Ergebnisse verfälscht.

Broadcast News Der Broadcast News Korpus ist ebenfalls ein englischer Datensatz und besteht aus Nachrichtensendungen, die über das Fernsehen oder das Radio aufgenommen wurden. Mit einem Vokabular von über 65.000 Wörtern ist er der größte Datensatz im Test.

4.2 Evaluierungsmaße

Konfidenz Fehlerrate Die Konfidenz Fehlerrate (eng.: *Confidence Error Rate*) (CER) ist relativ einfach definiert als die Anzahl der falsch markierten Wörter geteilt durch die gesamte Anzahl an Wörtern. Der Wert, der unter Verwendung eines Konfidenzmaßes erzielt wird, muss zur Bewertung mit der CER des reinen Erkennungssystems (*Baseline*) verglichen werden. Die Güte eines Maßes drückt sich in einer niedrigeren CER aus. Ist die CER mit und ohne Verwendung eines Konfidenzmaßes gleich, so stellt das Maß keine Informationen bereit, die nicht schon das Baseline-System besitzt.

Tabelle 2: Konfidenz Fehlerraten für die *Hypothesis Density* ($C_{D_{avg}}$) und die *Acoustic Stability* (C_A) auf den verschiedenen Datensätzen [14]. Die Skalierungsfaktoren für alle Verfahren wurden auf einem Teil des jeweiligen Datensatzes optimiert.

Datensatz	Baseline [%]	$C_{D_{avg}}$ [%]	C_A [%]
ARISE	13,6	11,7	8,2
Verbmobil'98	27,3	26,0	22,1
NAB 20k	11,3	11,3	9,9
NAB 64k	9,2	9,1	8,0
Broadc. News	27,7	27,7	25,8

Tabelle 3: Konfidenz Fehlerraten für die Verfahren auf Basis von N-Besten Listen auf den verschiedenen Datensätzen [14]. Die Skalierungsfaktoren für alle Verfahren wurden auf einem Teil des jeweiligen Datensatzes optimiert.

Datensatz	Baseline [%]	N-Besten Listen (C_N) [%]			
		$N = 100$	$N = 200$	$N = 300$	$N = 1000$
ARISE	13,6	8,9	8,6	9,0	9,1
Verbmobil'98	27,3	21,7	21,2	21,1	21,6
NAB 20k	11,3	9,4	9,2	9,1	9,2
NAB 64k	9,2	7,5	7,5	7,6	7,6
Broadc. News	27,7	25,3	25,3	25,2	25,4

Tabelle 4: Konfidenz Fehlerraten für die Verfahren auf Basis von Wortgraphen auf den verschiedenen Datensätzen [14].Die Skalierungsfaktoren für alle Verfahren wurden auf einem Teil des jeweiligen Datensatzes optimiert.

Datensatz	Baseline [%]	Wortgraphen (C_{W_*}) [%]			
		C_W	$C_{W_{sec}}$	$C_{W_{med}}$	$C_{W_{max}}$
ARISE	13,6	11,5	8,9	8,8	8,9
Verbmobil'98	27,3	23,3	19,0	20,0	18,9
NAB 20k	11,3	10,3	9,2	9,2	9,2
NAB 64k	9,2	8,4	7,2	7,2	7,2
Broadc. News	27,7	23,7	20,6	20,4	20,6

Statt der CER wird auch häufig die Konfidenzgenauigkeit (eng.: *Confidence Accuracy*) (CA) verwendet, die sich einfach durch $CA = 1 - CER$ definiert. Also die Anzahl der Taggingfehler geteilt durch die Anzahl der Wörter.

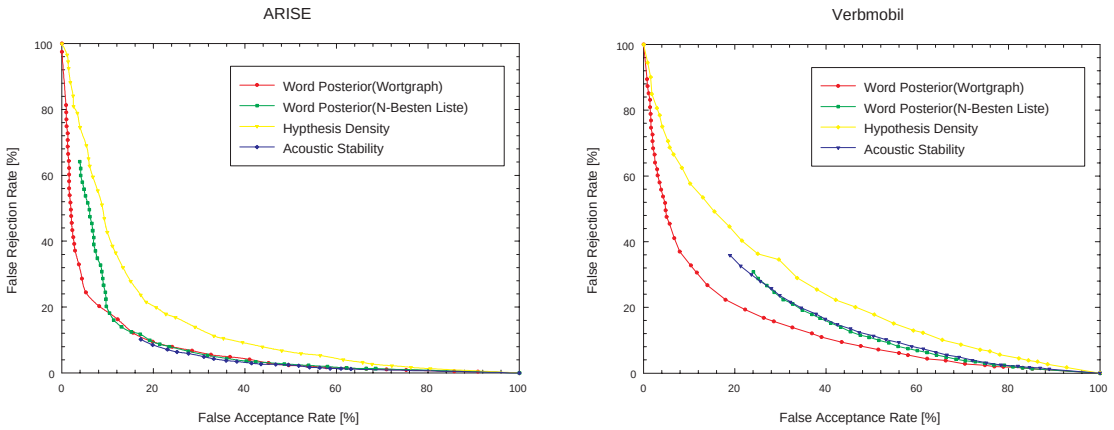
Detection Error Trade Off Bei der Einordnung eines Wortes in eine der beiden Klassen *Korrekt* oder *Falsch* können 2 Arten von Fehlern auftreten. Zum einen kann das Wort irrtümlicher Weise als korrekt markiert werden (*False Acceptance*), zum anderen kann das Wort als falsch markiert werden, obwohl es richtig erkannt wurde (*False Rejection*). Es ist leicht zu erkennen, dass es einen Trade-Off zwischen diesen beiden Werten gibt. Werden beide Werte gegeneinander in einem Diagramm aufgetragen, wird von einer *Detection Error Trade Off*-Kurve (DET) gesprochen. Diese Kurve ist ein gutes Bewertungskriterium für ein Konfidenzmaß, da anhand einer solchen Kurve beurteilt werden kann wie sich das System in bestimmten Situationen verhält. So kann ein Konfidenzmaß anhand einer DET-Kurve für das jeweilige System optimiert werden, in dem es verwendet werden soll. Ist eine hohe Sicherheit vonnöten, wird die False-Acceptance-Rate soweit optimiert, bis die False-Rejection-Rate am zumutbaren Maximum ist und umgekehrt.

Der Nachteil dieser Kurve ist allerdings das die Werte der False-Acceptance-Rate und der False-Rejection-Rate von der Güte des Verwendeten Erkennungssystems abhängt und zudem auch noch von der a-Priori Wahrscheinlichkeit für die beiden Klassen *Korrekt* und *Falsch*.

Equal Error Rate Die *Equal Error Rate* (EER) kann anhand der DET-Kurve für ein Konfidenzmaß abgelesen werden und bezeichnet den Wert bei dem die beiden Raten gleich groß sind. Somit kann abgeschätzt werden in wie weit sich das Konfidenzmaß für einen allgemeinen Einsatz eignet. Dieser Wert hat allerdings offensichtlicherweise den gleichen Nachteil wie die DET-Kurve.

4.3 Resultate

Wie in den Tabellen 2, 3 und 4 zu sehen ist verbessert sich die Erkennungsrate durch die Anwendung der Verfahren zumindest auf manchen Datensätzen. Es fällt auf, dass $C_{D_{avg}}$ (Hypothesis Density mit der durchschnittlichen Hypothesendichte für die betrachtete Worthypothese) immer die schlechteste Leistung bringt und manchmal sogar (Broadcast News, NAB 20k) keine Leistungssteigerung durch die Verwendung dieses Verfahrens erzielt werden kann. C_A (Acoustic Stability) und C_N (Word Posterior Wahrscheinlichkeiten auf N-



(a) (b)

Abbildung 9: DET-Kurven der vorgestellten Verfahren auf dem (a) ARISE- und (b) Verbmobil-Datensatz [14]. Zu beachten ist die gute Performance der Acoustic Stability (C_A) auf den ARISE-Daten.

Besten Listen) erbringen auf allen Datensätzen eine vergleichbare Performance, wobei C_A auf dem ARISE-Datensatz sogar die beste Leistung aller Verfahren zeigt. Durch die Verwendung von $C_{W_{max}}$ (Word Posterior Wahrscheinlichkeiten auf Wortgraphen mittels des Forward-Backward Algorithmus) als Konfidenzmaß können fast auf allen Testkörpern die besten Resultate erzielt werden.

Die DET-Kurven (Abbildungen 9 und 10) spiegeln dieses Bild wieder. $C_{W_{max}}$ stellt sich auf allen Datensätzen als das beste Verfahren dar. Die Überlegenheit von $C_{W_{max}}$ zeigt sich insbesondere auf dem schwierigen Broadcast News Korpus. Hier erbringen alle anderen Verfahren nur eine mittelmäßige Leistung, wohingegen $C_{W_{max}}$ eine wesentlich bessere Performance zeigt. Nur auf dem ARISE-Datensatz sind $C_{W_{max}}$, C_N und C_A ungefähr gleichwertig. Dies ist auf das mit ca. 1000 Wörtern relativ kleine Vokabular und die im Durchschnitt sehr kurzen Sätze[8] des ARISE-Datensatzes zurückzuführen. Allerdings ist bei diesen geringen Unterschieden ebenfalls $C_{W_{max}}$ zu favorisieren, da der Berechnungsaufwand für dieses Verfahren geringer ist als der Aufwand zur Berechnung der Acoustic-Stability.

Eine Möglichkeit den, durch das Berechnen der alternativen Satzhypothesen, hohen Aufwand der Acoustic-Stability zu senken besteht darin, die M besten Hypothesen zum Vergleich heranzuziehen, die bereits beim eigentlichen Erkennungsschritt ermittelt wurden und somit schon in einer N-Besten Liste oder einem Wortgraphen gespeichert sind. Dadurch fallen die weiteren Erkennungsschritte mit den unterschiedlichen Gewichtungen des Sprach- bzw. Akustischen Modells weg. Dieses leicht modifizierte Verfahren (C'_A) erzielt nur geringfügig schlechtere Ergebnisse als das Originalverfahren und kann wesentlich schneller berechnet werden.

Die Phonem-basierten Verfahren aus 2.3.1 zeigen in [11] keine bemerkenswerte Performance. Das in [3] vorgestellte Verfahren aus 2.3 ist von seiner Leistung ungefähr mit $C_{D_{avg}}$ gleichzusetzen, welches wie weiter oben erwähnt eine schlechtere Leistung bringt als die anderen hier näher betrachteten Verfahren.

[8] Die durchschnittliche Satzlänge in diesem Korpus beträgt nur 3 Wörter.

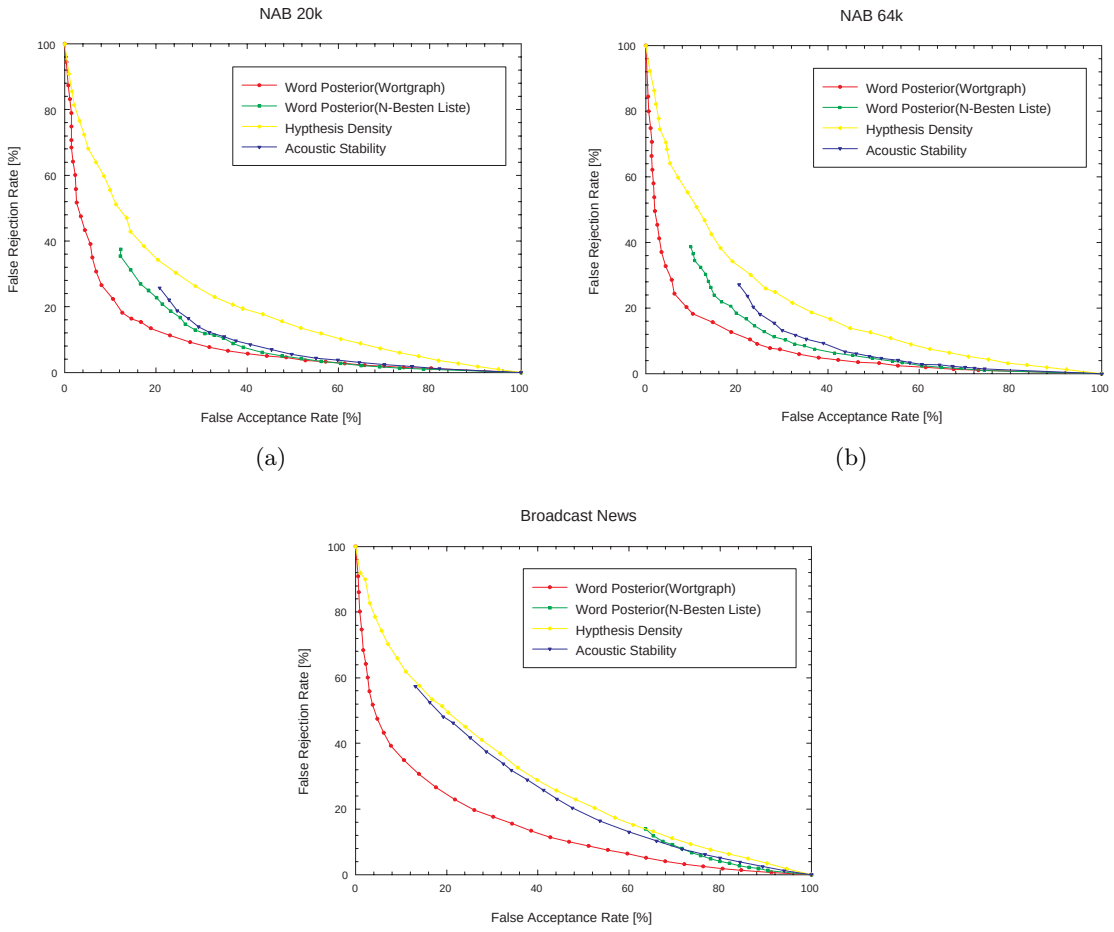


Abbildung 10: DET-Kurven auf dem (a) NAB 20k-, dem (b) NAB 64k und dem (c) Broadcast News-Datensatz [14]. Es ist zu erkennen, dass $C_{W_{max}}$ auf allen Daten die besten Leistungen erbringt. C_N und C_A sind auf diesen Testkörpern ungefähr gleichwertig.

Somit scheint $C_{W_{max}}$ zur Zeit eines der besten Konfidenzmaße im Bereich der Spracherkennung zu sein.

5 Anwendung von Konfidenzmaßen in der Praxis

Wie bereits eingangs (s. Abschnitt 1) erwähnt sind Dialogsysteme ein mögliches Gebiet zum Einsatz von Konfidenzmaßen. Bei der Optimierung der Benutzerinteraktion mit einem solchen Systems muss zum einen darauf geachtet werden, dass alle wichtigen Informationen zuverlässig ermittelt werden können, zum anderen dürfen die Nutzer eines Dialogsystems nicht durch unnötige oder verwirrende Verifikationsschritte abgeschreckt werden. Nach [2] hat sich gezeigt, dass eine implizite Verifikation von den Anwendern oft als verwirrend empfunden wird. So fühlen sich viele Menschen überfordert, wenn die implizite Verifikation mit der Frage nach einer zusätzlichen Information verbunden ist. Wenn ein Anrufer eines Dialogsystems zur Fahrplanauskunft beispielsweise durch „Ich möchte von Arnhem

nach Amsterdam fahren.“ Abfahrts- und Zielbahnhof bestimmt, das System allerdings statt „Arnhem“ „Haarlem“ versteht und Verifikation dieser Daten mit der Frage nach der Abfahrtszeit verknüpft, „Wann möchten Sie von Haarlem nach Amsterdam fahren?“, ist für den Nutzer nicht klar wie er sich verhalten soll.

S1	Van waar naar waar wilt u reizen?	*From where to where do you want to travel?*
U1	Ik zou graag naar Amsterdam reizen.	*I'd like to travel to Amsterdam.*
S2	Wilt u naar Amsterdam?	*Do you want to go to Amsterdam?*
U2	Ja, dat klopt.	*Yes, that's right.*
S3	Waar vandaan wilt u vertrekken?	*Where do you want to leave?*
U3	Uit Haarlem, alstublieft.	*From Haarlem, please.*
S4	Wilt u uit Arnhem vertekken?	*Do you want to leave from Arnhem?*
U4	Nee, Haarlem!	*No, Haarlem!*
...	...	...

Abbildung 11: Ein Dialog, bei dem das Dialogsystem die gemachten „Eingaben“ durch explizite Schritte (S2 und S4) verifiziert [2].

Im Rahmen des ARISE (Automatic Railways Information Systems in Europe) Projektes, auf das auch der entsprechende Datensatz aus Abschnitt 4.2 zurück geht, wird derzeit aus dem oben genannten Grund hauptsächlich die explizite Verifikation der Daten angewandt. Der Nachteil der expliziten Verifikation ist, dass sich ein Dialog durch die zusätzlichen Schritte zum Teil drastisch verlängern kann. Abbildung 11 zeigt einen Ausschnitt aus einem realen Anruf, der im holländischen ARISE-System eingegangen ist, wobei die Fragen des Systems mit S und die Antworten des Anrufers mit U bezeichnet werden.

Es ist leicht zu erkennen, dass dieser Dialog durch implizite Verifikation um knapp die Hälfte verkürzt werden kann (s. Abbildung 12). Aus diesem Grund wird versucht ein Konfidenzmaß in dieses System zu integrieren. Durch die Anwendung eines solchen Maßes kann zwischen den beiden Verifikationsmöglichkeiten gewählt werden. Ist sich das System sehr sicher ein Schlüsselwort (z.B. den Abfahrts- oder Zielbahnhof) richtig erkannt zu haben, braucht dieses nur noch implizit verifiziert zu werden. Bei unsicheren Wörtern hingegen, wird besser ein expliziter Verifizierungsschritt angewandt, um den Nutzer nicht zu verwirren.

Im weiteren planen die Autoren das in [17] vorgestellte Verfahren zur Konfidenzbestimmung in das Auskunftssystem zu integrieren.

S1	Van waar naar waar wilt u reizen?	*From where to where do you want to travel?*
U1	Ik zou graag naar Amsterdam reizen.	*I'd like to travel to Amsterdam.*
S2	Naar Amsterdam. Waar vandaan wilt u vertrekken?	*To Amsterdam. From where do you want to leave?*
...	...	...

Abbildung 12: Die Anwendung eines Konfidenzmaßes erlaubt Eingaben, die relativ sicher erkannt wurden nur noch implizit zu verifizieren. Dies geschieht hier durch einfaches Wiederholen der Eingaben [2].

6 Zusammenfassung und Ausblick

In dieser Seminararbeit wurde, nachdem der Autor in Abschnitt 1 die Verwendung von Konfidenzmaßen motiviert hatte, zunächst auf Konfidenzmaße im Allgemeinen eingegangen. Abschnitt 2 stellte die notwendigen Definitionen zur Verfügung, damit später in den Teilen 2.2, 2.3 und 3 genauer auf verschiedene Ansätze zur Berechnung von Konfidenzmaßen eingegangen werden konnte. Es wurden in dieser Arbeit insbesondere die Acoustic Stability (s. Abschnitt 2.2.1), die Hypothesis Density (s. Abschnitt 2.2.2) und die Verwendung und Berechnung von Word Posterior Wahrscheinlichkeiten (s. Abschnitt 2.3.2 und 3) als Konfidenzmaße vorgestellt und in Abschnitt 4.2 diskutiert. Dabei stellte sich heraus, dass zu Zeit die Word Posterior Wahrscheinlichkeiten als Konfidenzmaß alle anderen Verfahren an Leistung übertreffen.

Zukünftige Entwicklungen beschäftigen sich sehr wahrscheinlich zunächst mit der Optimierung der vorgestellten Verfahren und der Integration existierender Methoden in Spracherkennungssysteme (s. [2]). Desweiteren sind noch mehrere verbesserte Verfahren zu erwarten, da alle derzeitigen Konfidenzmaße noch weit von einer optimalen Lösung entfernt sind.

Literatur

[1] A. Asadi, R. Schwartz, and J. Makhoul. Automatic Detection of New Words in a Large-Vocabulary Continuous Speech Recognition System. In *Proceedings of the 15th International Conference on Acoustic, Speech and Signal Processing (ICASSP'90)*, pages 125–128. IEEE Computer Society Press, 1990.

[2] G. Bouwman, J. Sturm, and L. Boves. Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the ARISE Project. In *Proceedings of the 24st International Conference on Acoustic, Speech and Signal Processing (ICASSP'99)*, pages 493–496. IEEE Computer Society Press, 1999.

[3] S. Cox and R. Rose. Confidence Measures for the Switchboard Database. In *Proceedings of the 21st International Conference on Acoustic, Speech and Signal Processing (ICASSP'96)*, pages 511–514. IEEE Computer Society Press, 1996.

[4] M. Finke and T. Zeppenfeld. LVCSR Switchboard April 1996 Evaluation Report. In *Proceedings of the LVCSR Hub 5 Workshop*, 1996.

[5] A. Kellner, B. Rueber, F. Seide, and B.H. Tran. PADIS – an Automatic Telephone Switchboard and Directory Information System. *Speech Communication*, 23:95–111, 1997.

[6] T. Kemp and T. Schaaf. Estimating Confidence Using Word Lattices. In *Proceedings of the 4th European Conference on Speech, Communication and Technology (EUROSPEECH'97)*, pages 827–830, 1997.

[7] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[8] H. Ney. Acoustic Modeling of Phoneme Units for Continuous Speech Recognition. In *Proceedings of the 5th European Signal Processing Conference (EUSIPCO'90)*, pages 65–72. Elsevier Science Publishers, 1990.

[9] H. Qiu. Confidence Measure for Speech Recognition Systems. Master Thesis, 1996.

[10] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[11] Z. Rivlin, M. Cohen, V. Abrash, and T. Chung. A Phone-Dependent Confidence Measure for Utterance Rejection. In *Proceedings of the 21st International Conference on Acoustic, Speech and Signal Processing (ICASSP'96)*, pages 515–517. IEEE Computer Society Press, 1996.

[12] T. Schaaf and T. Kemp. Confidence Measures for Spontaneous Speech Recognition. In *Proceedings of the 22nd International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 875–878. IEEE Computer Society Press, 1997.

[13] A. J. Viterbi. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[14] F. Wessel. Word Posterior Probabilities for Large Vocabulary Continuous Speech Recognition. Dissertation, 2002.

[15] F. Wessel, K. Macherey, and H. Ney. A Comparison of Word Graph and N-Best List Based Confidence Measures. In *Proceedings of the 6th European Conference on Speech, Communication and Technology (EUROSPEECH'99)*, pages 315–318, 1999.

[16] F. Wessel, K. Macherey, and R. Schlüter. Using Word Probabilities as Confidence Measures. In *Proceedings of the 23rd International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)*, pages 225–228. IEEE Computer Society Press, 1998.

[17] F. Wessel, R. Schlüter, K. Macherey, and H. Ney. Confidence Measures for Large Vocabulary Continuous Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3):288–298, 2001.

[18] D. Willett, C. Neukirchen, and G. Rigoll. Efficient Search with Posterior Probability Estimates in HMM-based Speech Recognition. In *Proceedings of the 23rd International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)*, pages 821–824. IEEE Computer Society Press, 1998.

[19] D. Willett, A. Worm, C. Neukirchen, and G. Rigoll. Confidence Measures for HMM-Based Speech Recognition. In *Proceedings of the 5th International Conference on Spoken Language Processsing (ICSLP'98)*, pages 3241–3244. Australian Speech, Science and Technology Association, Incorporated (ASSTA), 1998.