# STIMMAKTIVIERUNG EINES SPRACHERKENNERS

*Marco Kühne und Matthias Wolff*

*Institut für Akustik und Sprachkommunikation, TU Dresden*
*Marco.Kuehne\Matthias.Wolff@ias.et.tu-dresden.de*

**Abstract:** Der vorliegende Artikel befasst sich mit der Stimmaktivierung eines Spracherkenners per Schlüsselwort. Als primäre Zielsetzung wurde die Erkennung des Schlüsselwortes bei möglichst geringer Fehlalarmrate verfolgt. Prosodische Faktoren wie die Stimmmelodie fanden bisher in den Verfahren zur Stimmaktivierung nur sehr geringe Beachtung. Der Untersuchungsschwerpunkt lag daher auf der Entwicklung eines Verfahrens zur Integration prosodischer Merkmale in die Erkennungsstrategie.

## 1 Einleitung

Einen kritischen Punkt für den praktischen Einsatz von Spracherkennern stellt die Aktivierung des Systems dar. Viele aktuelle Systeme fordern eine „push-to-talk“ Aktivierung des Systems durch den Benutzer. Für die meisten Anwendungen ist eine manuelle Aktivierung jedoch unzweckmäßig. Eine automatische Stimmaktivierung des Spracherkenners ermöglicht oftmals eine deutlich ergonomischere Bedienung durch ein komplettes „hands- and eyes-free“ Sprachinterface. Üblicherweise werden solche Systeme als Schlüsselwortdetektoren implementiert, welche den Spracheingabestrom kontinuierlich nach einem vordefinierten Schlüsselwort durchsuchen [1]. Zur Modellierung des Schlüsselwortes werden speziell trainierte Wort-HMM's benutzt. Alle restlichen Sprachsegmente, welche kein Schlüsselwort darstellen sowie alle nichtsprachlichen Signalabschnitte werden durch Füller- und Müllmodelle repräsentiert. Die größte Herausforderung besteht in der robusten Rückweisung so genannter „Out-Of-Vocabulary“ (OOV) Wörter, um Fehlalarme zu vermeiden [2]. Doch selbst im Falle eines idealen Schlüsselwortdetektors können Fehlalarme auftreten. Wird das Schlüsselwort in einem anderen Kontext als in einem Kommando artikuliert, führt dies unweigerlich zu einer fehlerhaften Aktivierung des Systems. Beispielsweise kann das Aktivierungswort „Computer“ auch in normaler Unterhaltungssprache auftauchen, wodurch das System fälschlicherweise aktiviert werden könnte. Aus diesem Grund werden im praktischen Einsatz oft exotische (benutzerunfreundliche) Schlüsselwörter verwendet. Die in dieser Arbeit durchgeführten Experimente haben gezeigt, dass Versuchspersonen das Schlüsselwort beim Ansprechen des Spracherkenners besonders hervorheben. Damit stehen prosodische Faktoren wie die Grundfrequenz (Stimmmelodie), die Intensität sowie Silben- bzw. Wortdauer zur Verfügung, um zu bestimmen, ob das Schlüsselwort Teil eines Kommandos war oder nicht. In aktuellen Verfahren zur Stimmaktivierung blieben prosodische Faktoren wie die Stimmmelodie und die Intensität weitestgehend unberücksichtigt. YAMASHITA und MIZOGUCHI haben gezeigt, dass die Erkennungsleistung eines Schlüsselwortdetektors von der Einbeziehung prosodischer Wissensquellen profitieren kann [3]. Anhand eines sprecherabhängigen DTW (Dynamic Time Warping) Ansatzes wird die Grundfrequenzkontur einer Schlüsselworthypothese mit einem abgelegten Referenzmuster verglichen. Für ein sprecherunabhängiges System ist die Wahl nur eines einzigen Referenzmuster kritisch zu betrachten. Wir schlagen daher vor, für die Modellierung der prosodischen Muster des Schlüsselwortes ein HMM zu benutzen, welches die Schlüsselworthypothesen eines phonetischen Spra-

cherkenners anhand prosodischer Merkmale auf ein Ansprechen des Spracherkenners verifiziert. Der Untersuchungsschwerpunkt lag daher auf der Entwicklung eines robusten Verfahrens zur Stimmaktivierung durch die Integration prosodischer Merkmale in die Erkennungsstrategie.

## 2 Das prosodische Modell

Bereits in [4] haben LJOLJE und FALLSIDE gezeigt, dass HMM's in der Lage sind, prosodische Muster wie die Grundfrequenz und Intensität zu beschreiben. Zur Modellierung der prosodischen Eigenschaften des Schlüsselwortes beim Ansprechen des Spracherkenners wurde ein Hidden Markov Modell erstellt und trainiert. Als Merkmale zur Beschreibung der prosodischen Muster wurden die Grundfrequenz $F_0$, die Root-Mean-Square (RMS) Energie sowie deren erste zeitliche Ableitungen verwendet. Die Berechnung der beiden Größen $F_0$ und $RMS$ Energie erfolgte durch das ESPS-Programm get_f0. Alle Beobachtungsvektoren wurden im Abstand von 10 ms berechnet. Stimmhafte Segmente mit weniger als 30 ms wurden auf einen Wert von Null (stimmlos) zurückgesetzt. Stimmlose Bereiche wurden durch eine Spline-Funktion interpoliert. Für eine sprecherunabhängige Weiterverarbeitung wurden die $F_0$ und $RMS$ Konturen normiert, geglättet und auf eine logarithmische Skala konvertiert. Abbildung 1 zeigt gemittelte $F_0$-Konturen des Schlüsselwortes „Computer" für 30 Sprecher. Konturen des Schlüsselwortes, welche innerhalb eines Kommandos artikuliert wurden, sind grau dargestellt. Die schwarzen Konturen repräsentieren Schlüsselwörter welche aus normaler Sprache stammen. Der Vergleich zeigt, dass die Stimmmelodie deutliche Unterschiede zwischen Kommando- und normaler Sprache aufweist.

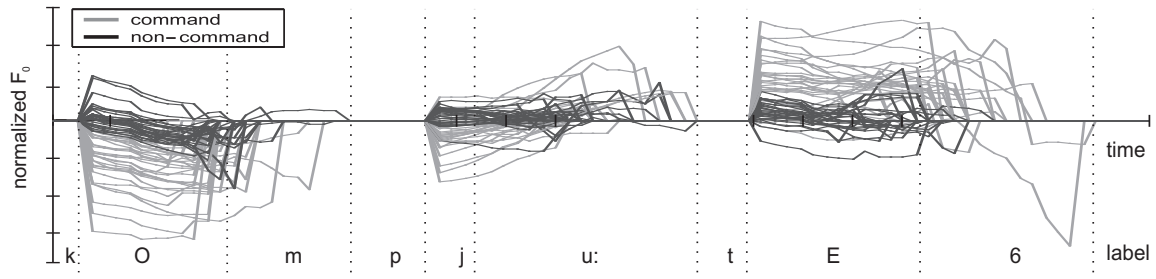


**Abbildung 1** - $F_0$ Konturen des Schlüsselwortes „Computer" aus Kommandophrasen (command) und normaler Sprache (non-command)

Zur Modellierung der prosodischen Merkmale wurde ein prosodisches HMM

$$\text{PHMM} = \left(V, E, \{\mathcal{N}\}, \nu^{(V)}, \pi^{(E)}\right) \tag{1}$$

bestehend aus einer Knotenmenge $V$ und einer gerichteten Kantenmenge $E \subseteq V \times V$ erstellt. Durch die Abbildung $\nu^{(V)} : V \rightarrow \{\mathcal{N}\}$ wird jedem Knoten eine 4-dimensionale normalverteilte Wahrscheinlichkeitsdichtefunktion $\mathcal{N}_i(\underline{\mu}_i, \Sigma_i) \in \{\mathcal{N}\}$ mit vollbesetzten Kovarianzmatrizen $\Sigma_k$ zugeordnet. Jede Kante des Graphen trägt eine Übergangswahrscheinlichkeit $\pi^{(E)} : E \rightarrow (0, 1]$. Das Modell wurde anhand eines VITERBI-Verfahrens mit 138 Beispielen von 3 männlichen und 3 weiblichen deutschen Sprechern trainiert. Die Trainingsdaten wurden separat mit

den jeweiligen Sprechern aufgenommen und manuell zurechtgeschnitten. Ein Großteil der Trainingsbeispiele wurde visuell auf einen korrekten Grundfrequenzverlauf geprüft und bei fehlerhafter F0-Extraktion entsprechend ausselektiert. Anhand des VITERBI-Algorithmus wurden alle Trainingsbeispiele über mehrere Iterationen hinweg durchlaufen, wobei jeweils die akustischen Modellparameter (Mittelwertvektoren $\underline{\mu}_i$ und Kovarianzmatrizen $\Sigma_i$) neu geschätzt wurden. Als Initialschätzung wurden die Merkmalvektoren gleichmäßig auf die Zustände des HMM's verteilt. Nach einer festgelegten Anzahl von Iterationen wurde der Trainingsprozess beendet.

Während der Erkennungsphase wird zunächst eine einfache linear verkettete Graphenstruktur $\mathcal{X}$ aufgebaut, welche jedem Merkmalvektor genau einen Knoten zuordnet. Anschließend wird dieser Graph auf die Modellstruktur PHMM mittels einer VITERBI-Suche abgebildet:

$$\mathcal{U}^* = \underset{\mathcal{U} \subseteq \text{PHMM} \times \mathcal{X}}{\operatorname{argmax}} \left[ \sum_{i=1}^{|\mathcal{U}|} D_{LL}(\underline{x}(u_i) | \mathcal{N}(u_i)) \right] \tag{2}$$

Dabei bezeichnet $\mathcal{U}^*$ denjenigen Pfad durch den Modellgraphen PHMM, welcher die Summe der logarithmierten Emissionsdichtewerte des Merkmalvektors $\underline{x}$ maximiert. Als lokale Bewertung $D_{LL}$ für die Beobachtung des Merkmalvektors $\underline{x}$ in einem bestimmten Zustand $k$ wird der logarithmierte Emissionsdichtewert (Log-Likelihood) verwendet:

$$D_{LL}(\underline{x} | \mathcal{N}(v_k)) = -(\underline{x} - \underline{\mu}_k)^T \Sigma_k^{-1} (\underline{x} - \underline{\mu}_k) - \ln |\Sigma_k| \tag{3}$$

Mit $LL^*(\underline{X}|\text{PHMM})$ wird die Log-Likelihood Summe der Emissionswerte entlang des optimalen Pfades $\mathcal{U}^*$ bezeichnet.

Für eine Modellierung der Zustandsdauern wurde eine heuristische Methode basierend auf normierten Zustandsdauerhistogrammen implementiert. Während der Trainingsphase werden Histogramme der Zustandsdauern für jeden Modellzustand erstellt. Diese Histogramme werden benutzt, um Wahrscheinlichkeitsdichtefunktionen der Zustandsdauern als Mischung von Normalverteilungsdichten zu schätzen. In der Erkennungsphase werden die normierten Zustandsdauern $\tau_i$ nach erfolgter VITERBI-Segmentierung berechnet und mit den im Training geschätzten Zustandsdauerverteilungen $p(\tau_i)$ verglichen. Die durch die Überprüfung der Zustandsdauern modifizierte Log-Likelihood Summe wird direkt als prosodisches Konfidenzmaß

$$PS = LL^*(\underline{X}|\text{PHMM}) + \eta \sum_{j=1}^{N} \log\left[p_j(\tau_j)\right] \tag{4}$$

benutzt. Das Konfidenzmaß $PS$ kann somit als Indikator dafür angesehen werden, wie ähnlich die Beobachtungen $\underline{X}$ und das Modell PHMM sind.

## 3 Kommandoerkennung mit Stimmaktivierung

Für die durchgeführten Experimente wurde das prosodischen Modell PHMM in den UASR-Spracherkenner integriert (siehe Abbildung 2). Das Training der akustischen SMG-Modelle erfolgte mit etwa 30 Stunden Sprachmaterial des VERBMOBIL-Korpus [6]. Eine detaillierte Beschreibung des UASR-Spracherkenners ist in [9] zu finden. Der Erkenner wurde für Kommandophrasen, bestehend aus dem Schlüsselwort gefolgt von einem Kommando, konfiguriert. Nach

erfolgter akustischer Erkennung werden die Schlüsselworthypothesen des Erkenners durch das prosodische Modell PHMM auf ein Ansprechen verifiziert.

Der Kommandoerkenner berechnet zwei Konfidenzmaße RPMS und RASD zur Überprüfung der Erkennungsergebnisse. Beide benutzen eine freie Phonemerkennung als Referenzbewertung. Dabei benutzen die Kommando- als auch die freie Phonemerkennung identische SMG-Monophonmodelle. Das erste Konfidenzmaß (**R**elative **P**honetic **M**atch **S**core)

$$RPMS = \frac{n_{match}}{n_{ref}} \tag{5}$$

berechnet auf Merkmalvektorebene die Anzahl übereinstimmender Phonemsymbole zwischen Kommando- und Referenzerkennung. Dabei bezeichnet $n_{match}$ die Anzahl der übereinstimmenden Symbole und $n_{ref}$ die Gesamtzahl an Phonemsymbolen in der Referenzerkennung. Je höher der RPMS Wert, desto höher die Konfidenz des Erkennungsergebnis. Die zweite Konfidenzgröße RASD (**R**elative **A**coustic **S**core **D**ifference) kombiniert die akustischen und prosodischen Log-Likelihood Bewertungen:

$$RASD = \left| \frac{(AS_{cmd} - AS_{ref}) - w \cdot (|PS| - |\overline{PS}|)}{AS_{cmd}} \right| \tag{6}$$

Dabei bezeichnet $AS_{cmd}$ die akustische Bewertung (Log-Likelihood) der durch das Lexikon beschränkten Kommandoerkennung und $AS_{ref}$ die Bewertung der unbeschränkten Referenzerkennung. Der zweite Term im Zähler integriert eine prosodische Bewertung in die Konfidenzberechnung, welche durch den Wichtungsfaktor $w$ skaliert werden kann. $\overline{PS}$ steht für die durchschnittliche prosodische Bewertung aller Schlüsselwörter [3]. Je kleiner die RASD Bewertung, desto höher die Konfidenz des Erkennungsergebnis. Durch den Wichtungsfaktor $w$ kann der Einfluss der prosodischen Verifikation durch das PHMM variiert werden. Ein Gewicht $w = 0$ führt auf ein Baseline-System (BL) ohne prosodische Informationen. Im Gegensatz dazu bezieht das System VA mit einem Gewicht $w > 0$ zusätzliche prosodische Wissensquellen in die Konfidenzberechnung ein.

Anhand der beiden Konfidenzmaße RASD und RPMS wird das Erkennungsergebnis bei zu niedriger Konfidenz zurückgewiesen. Für die Rückweisungslogik wurden zwei unterschiedliche Strategien getestet. Die erste (OP) verwendet zwei Schwellwerte $\alpha$ und $\beta$. Durch die beiden Schwellwerte wird ein fester Arbeitspunkt $OP(\alpha, \beta)$ des Erkenners eingestellt. Ein Erkennungsergebnis wird nur dann akzeptiert, wenn $RASD \leq \alpha$ and $RPMS \geq \beta$. Die zweite Rückweisungsstrategie (CMC) beruht auf einer Abstandsklassifikation der beiden Konfidenzgrößen. Dazu wird angenommen, dass sich die Verteilungen der Konfidenzgrößen für die Klassen $k \in \{cmd, oov\}$ als zweidimensionale Normalverteilungen mit Mittelwertvektor $\vec{\mu}_k$ und Kovarianzmatrix $\Sigma_k$ beschreiben lassen. Der Vektor $\vec{c} = (RASD\ RPMS)^T$ repräsentiert beide Konfidenzwerte eines Erkennungsergebnisses. Basierend auf der MAHALANOBIS-Distanz

$$d_k(\vec{c}) = (\vec{c} - \vec{\mu}_k)^T \Sigma_k^{-1} (\vec{c} - \vec{\mu}_k) \tag{7}$$

wird ein Erkennungsergebnis zurückgewiesen, wenn $d_{cmd}(\vec{c}) > d_{oov}(\vec{c})$.

## 4 Evaluierung der Erkennungsleistung

Die Evaluierung der Erkennungsleistung wurde anhand einer sprachgesteuerten PowerPoint-Präsentation zum Thema „Zeitreise durch die Computergeschichte“ durchgeführt. Dazu wurde

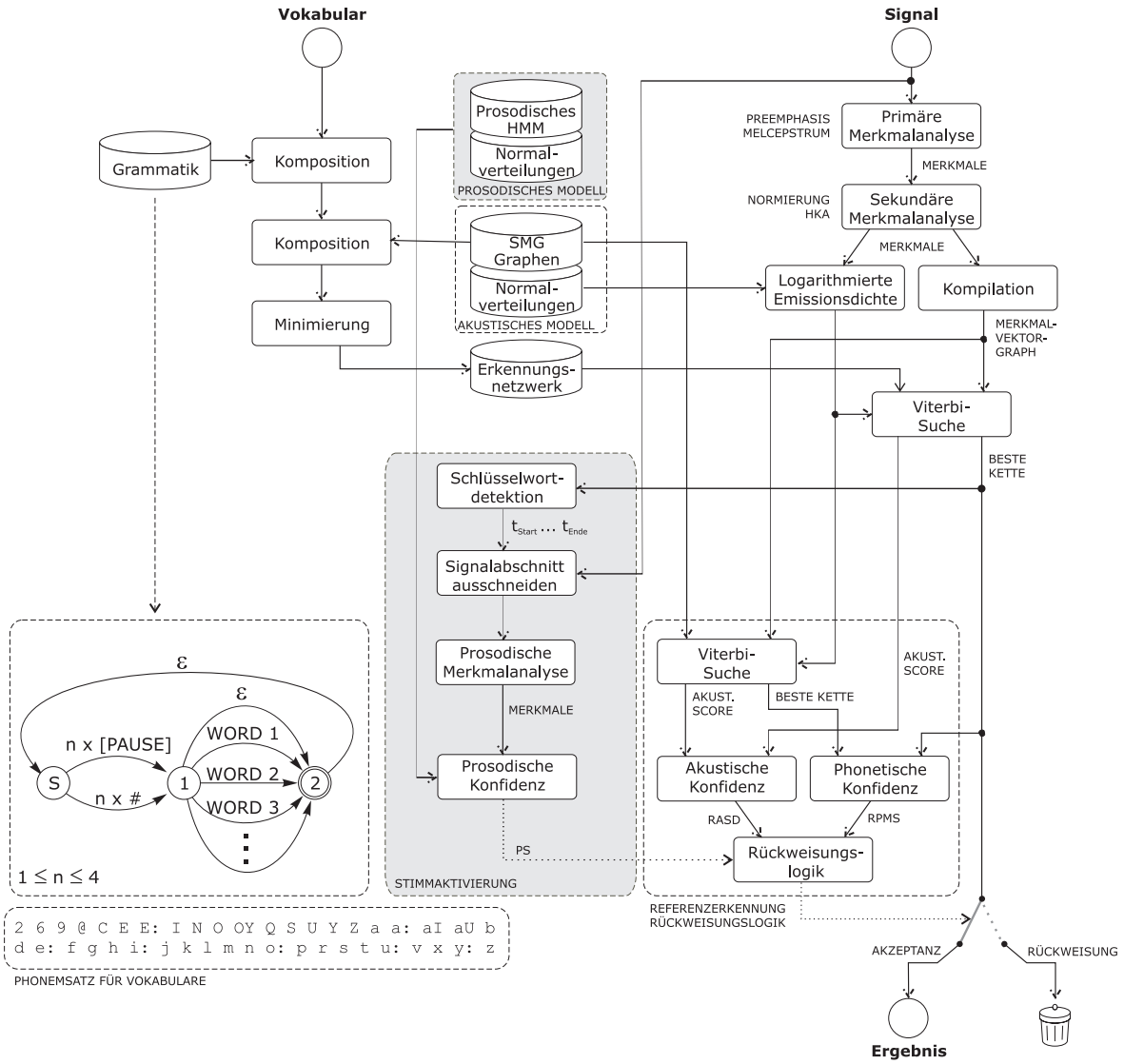


**Abbildung 2** - Integration der Stimmaktivierung in die UASR-Architektur

das gesamte Sprachmaterial von 24 männlichen und 6 weiblichen deutschen Sprechern mit einer Abtastfrequenz von 16 kHz, 16 Bit Quantisierung und mittels eines Headset-Mikrophons unter Laborbedingungen aufgezeichnet. Das Sprachmaterial enthielt 344 Phrasen mit dem Schlüsselwort „Computer“ gefolgt von einem Kommando zur Steuerung der Präsentation (z.B.: „nächste Folie“). Ferner waren 193 Phrasen enthalten, welche ebenfalls das Schlüsselwort beinhalteten, aber Teil des Vortrages waren und daher nicht innerhalb einer Kommandophrase artikuliert wurden. Das restliche Sprachmaterial wurde automatisch in 3630 OOV Phrasen segmentiert, deren Länge in etwa der einer Kommandophrase entsprachen. Die gesamte Datenbank umfasste 0,69 Stunden Sprache. Keiner der Sprecher wurde angewiesen, einen bestimmten Sprechstil zum Ansprechen des Erkenners zu verwenden. Das Vokabular zur Steuerung der Präsentation umfasste 7 Wörter. Basierend auf diesem Lexikon wurden vier verschiedene Kommandos durch eine explizite Grammatik definiert. Die durchgeführten Experimente wurden in drei verschiedene Aufgaben unterteilt. Zunächst wurde der Einfluss des prosodischen Wichtungsfaktors $w$ (siehe Gleichung 6) auf die Erkennungsleistung des Kommandoerkenners untersucht. Anschließend wurde evaluiert, inwieweit das prosodische Modell in der Lage ist zwischen Schlüsselwörtern in Kommandos und normaler Sprache zu unterscheiden. Diese Ergebnisse des PHMM wurden den

Resultaten von menschlichen Versuchspersonen in zwei Hörexperimenten gegenübergestellt. In der letzten Aufgabe wurde die Performanz des Kommandoerkenners unter dem Gesichtspunkt der Rückweisung von OOV Äußerungen evaluiert.

### 4.1 Einfluss des prosodischen Gewichts

Um den Einfluss des prosodischen Gewichts w auf die Erkennungsleistung zu messen, wurden die Kommandoerkennungsrate und die Fehlalarmrate für verschiedene Gewichte $w$ in einem festen Arbeitspunkt $OP_1(0,09; 0,25)$ untersucht (siehe Abbildung 3). Das Baseline-System (BL, $w = 0$) erreicht dabei eine Erkennungsrate von 95,1 % bei einer Fehlalarmrate von 28,5 % auf dem oben beschriebenen Testdaten. Durch schrittweise Erhöhung des prosodischen Gewichts wird die Fehlalarmrate signifikant verringert, während die Erkennungsrate nur leicht abfällt. Für ein Gewicht $w = 0,2$ erreicht das System VA eine Erkennungsrate von 90,1 % bei einer Fehlalarmrate von 4,5 %. Für die weiteren Experimente wurde im folgenden ein Gewicht $w = 0,2$ verwendet.

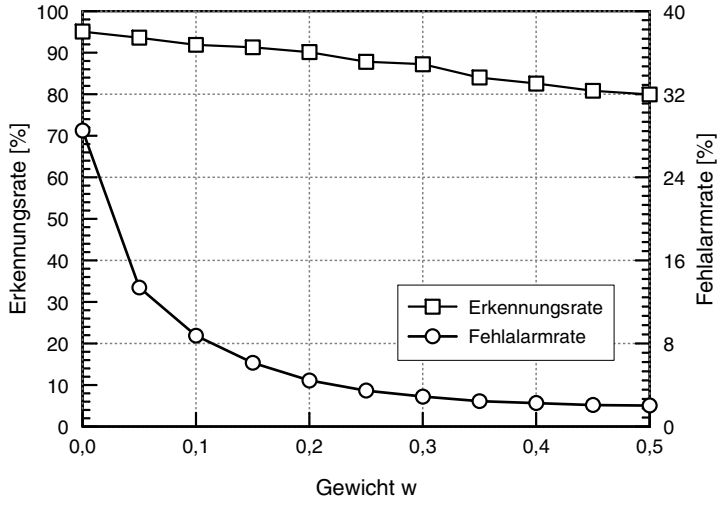


**Abbildung 3** - Einfluss des Wichtungsfaktors $w$ auf die Erkennungs- und Fehlalarmrate

### 4.2 Verwendung von Aussprachevarianten

Weiterhin wurde untersucht, ob mit Hilfe von Aussprachevarianten die Erkennung des Schlüsselworts „Computer“ verbessert werden kann. Abbildung 4 zeigt die Fehlalarmrate aufgetragen über der Fehlrückweisungsrate für den oben beschriebenen Kommandowortschatz aus 4 Kommandophrasen (jeweils inklusive Schlüsselwort). Im Test waren insgesamt 4167 Äußerungen von 30 verschiedenen Sprechern, davon 344 gültige Kommandos. Das Schlüsselwort „Computer“ wurde *a)* ohne Aussprachevarianten (kanonische Form) sowie *b)* mit 52 Aussprachvarianten modelliert. Tabelle 1 zeigt die Equal Error Rates. Für alle getesteten Konfigurationen konnten die Fehlerraten durch Verwendung von Aussprachevarianten für das Schlüsselwort leicht gesenkt werden.

### 4.3 PHMM vs. Menschliche Erkennungsleistung

In diesem Experiment wurden 50 isolierte Beispiele des Schlüsselwortes „Computer“ der Testdaten automatisch durch das PHMM als „normale Sprache“ oder „Ansprechen des Erkenners“

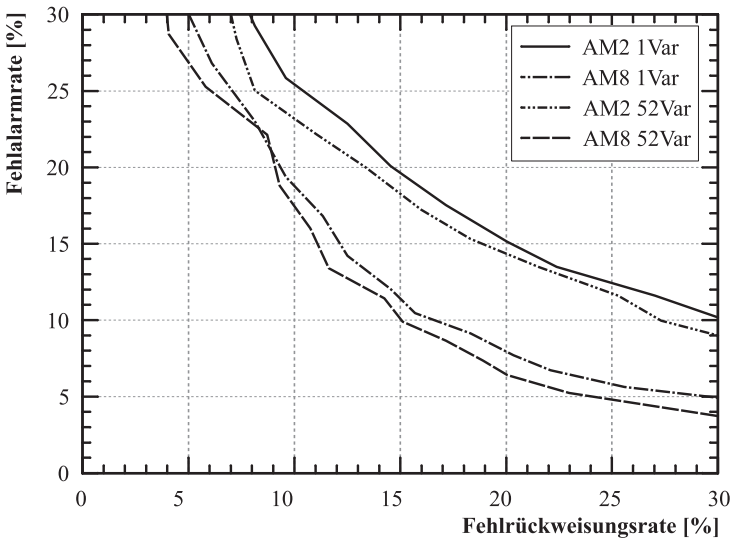


**Abbildung 4** - Phonetische Erkennungsleistung eines Stimmaktivierungssystems mit und ohne Aussprachevarianten für das Schlüsselwort „Computer“. AM2 und AM8 stehen für zwei verschiedene getestete akustische Modelle

**Tabelle 1** - Vergleich der Equal Error Rate für die akustischen Modelle AM2 und AM8 bei unterschiedlicher Anzahl an Aussprachevarianten für das Schlüsselwort

| | Equal Error Rate (EER) [%] | |
|---|---|---|
| | Schlüsselwort+Kommando | nur Schlüsselwort |
| AM2 1Var | 17,5 | / |
| AM2 52Var | 16,9 | / |
| AM8 1Var | 12,8 | 36,9 |
| AM8 52Var | 12,5 | 35,8 |

klassifiziert. Zusätzlich wurden 32 Versuchspersonen gebeten in zwei Hörexperimenten die identische Klassifikationsaufgabe vorzunehmen. Da die Hörbeispiele ausschließlich das Schlüsselwort beinhalteten, wurden die Testpersonen gezwungen ihre Entscheidung allein auf der Basis prosodischer Kriterien zu treffen. Die Beispiele des Hörtest wurden von 4 männlichen und 4 weiblichen Sprechern gesprochen. Im ersten Hörexperiment (T1) wurden die Testbeispiele über alle 8 Sprecher hinweg zufällig angeboten. Im zweiten Teil (T2) wurden die Hörproben nach Sprechern geordnet, sodass die Testbeispiele eines Sprechers nacheinander in zufälliger Reihenfolge angeboten wurden. Um die Versuchspersonen mit Informationen über den Sprechstil der einzelnen Sprecher zu unterstützen, bestand in T2 zusätzlich die Möglichkeit sich einen Beispielsatz des jeweiligen Sprechers anzuhören. Tabelle 2 vergleicht die durchschnittlichen Erkennungs- und Fehlalarmraten der Versuchspersonen mit der automatischen Klassifikation durch das PHMM. Im ersten Versuchsteil ist die Erkennungsleistung der Testpersonen nur geringfügig besser als die automatische Klassifikation. Dies deutet darauf hin, das die wesentlichsten prosodischen Muster durch das PHMM erfasst werden. Die Resultate des zweiten Tests T2 mit veränderten Versuchsbedingungen zeigen weiterhin die gute Fähigkeit des Menschen, sehr schnell auf unterschiedliche Sprechstile adaptieren zu können. Daher können Sprecheradapti-

onstechniken in Zukunft dazu beitragen, die prosodische Modellierung durch das PHMM weiter zu verfeinern.

| | Versuchspersonen | | PHMM |
|---|---|---|---|
| | T1 | T2 | Model |
| Erkennungsrate [%] | 81,1 | 88,3 | 74,2 |
| Fehlalarmrate [%] | 29,6 | 19,3 | 25,5 |

**Tabelle 2** - Gegenüberstellung der Resultate der Hörexperimente T1 und T2 und der automatischen Klassifikation durch das PHMM

### 4.4 Kommandoerkennung

Um die Leistungsfähigkeit der zusätzlichen prosodischen Informationen zu evaluieren, wurden die beiden Systeme BL und VA in ihrer Erkennungsleistung miteinander verglichen. Abbildung 5 zeigt die Performance beider Systeme für verschiedene Arbeitspunkteinstellungen des Erkenners. Gemessen wurden jeweils die Fehlalarm- und Fehlrückweisungsrate für die Kommandoerkennung. Das Baseline-System (BL) ohne prosodische Informationen erreicht bei einer Fehlrückweisungsrate von 10 % eine Fehlalarmrate von etwa 14 %. Bei der gleichen Fehlrückweisungsrate von 10 % erreicht das System VA mit prosodischer Information eine deutlich geringere Fehlalarmrate von 4 %. Zusammenfassend konnte die Equal Error Rate (EER) des Baseline-Systems von 13,1 % auf 6,9 % gesenkt werden. Zur Veranschaulichung der zwei-

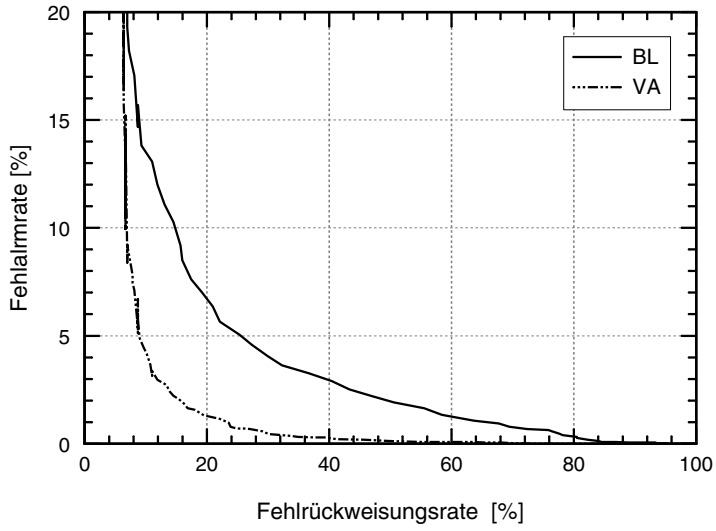


**Abbildung 5** - Vergleich Fehlalarm- und Fehlrückweisungsrate für die Systeme BL und VA

ten Rückweisungsstrategie zeigt Abbildung 6 die Verteilungen der beiden Konfidenzgrößen mit und ohne prosodische Verifikation. Zusätzlich sind die MAHALANOBIS-Trennfunktion im Kontrast zu den beiden Rückweisungsschwellen des festen Arbeitspunktes $OP_1$ eingetragen. Basierend auf der MAHALANOBIS-Abstandsklassifikation erreicht das Baseline-System eine Erkennungsrate von 89,0 % bei einer Fehlalarmrate von 6,1 %. Beide Größen konnten durch die zusätzliche prosodische Verifikation der Schlüsselworthypothesen deutlich verbessert werden. Das VA System resultiert in bei einer Erkennungsrate von 92,2 % bei einer Fehlalarmrate

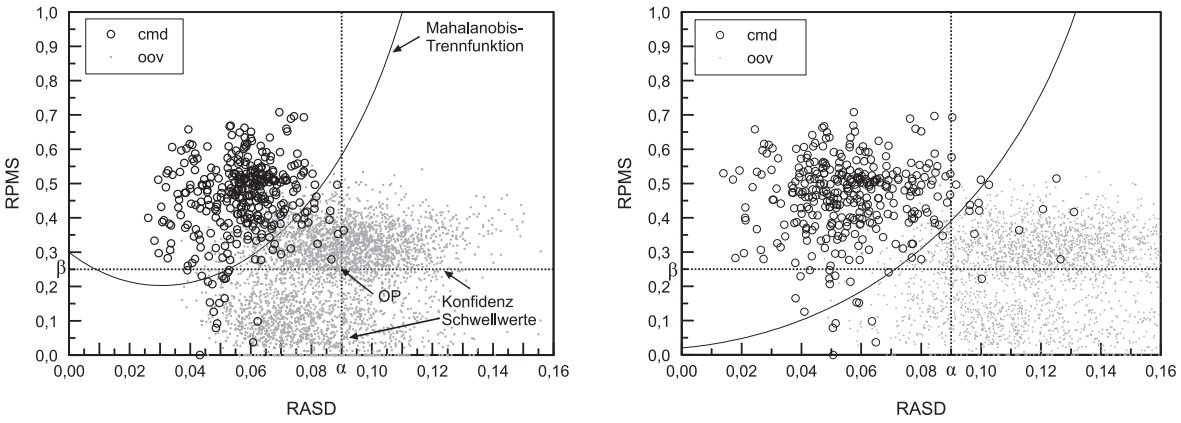


**Abbildung 6** - Verteilung der Konfidenzmaße RASD und RPMS für ein prosodisches Gewicht $w = 0$ (BL, links) und $w = 0,2$ (VA, rechts)

von 2,7 %. Damit können basierend auf repräsentativen Trainingsstichproben geeignete Trennfunktionen zur Rückweisung von OOV Phrasen geschätzt werden. Tabelle 3 fasst die Resultate der Erkennungsergebnisse für die Kommandoerkennung beider Systeme BL und VA für beide Rückweisungsstrategien zusammen.

| | $BL_{OP}$ | $VA_{OP}$ | $BL_{CMC}$ | $VA_{CMC}$ |
|---|---|---|---|---|
| Erkennungsrate [%] | 85,2 | 91,4 | 89,0 | 92,2 |
| Fehlalarmrate [%] | 13,1 | 6,9 | 6,1 | 2,7 |
| Fehlrückweisungsrate [%] | 13,1 | 6,9 | 10,3 | 7,7 |

**Tabelle 3** - Zusammenfassung der Erkennungsleistung für beide Systeme VA und BL abhängig von der verwendeten Rückweisungsstrategie OP bzw. CMC

## 5 Zusammenfassung und Ausblick

Die durchgeführten Experimente haben gezeigt, dass das Schlüsselwort beim Ansprechen des Spracherkenners besonders durch prosodische Faktoren hervorgehoben wird. Basierend auf dem Grundfrequenz- und Intensitätsverlauf des Schlüsselwortes „Computer“ wurde ein Hidden Markov Modell trainiert und in die Erkennungsstrategie eines Kommandoerkenners implementiert. Wir haben in Erkennungsexperimenten gezeigt, dass durch den Einsatz dieser prosodischen Wissensquellen die Erkennungsleistung eines herkömmlichen Erkenners signifikant gesteigert werden kann. Insbesondere die Anzahl der Fehlalarme konnte mit dem vorgestellten Ansatz deutlich gesenkt werden. In unserer weiteren Arbeit werden wir eine auf die Stimmaktivierung zugeschnittene Pausendetektion in das Verfahren integrieren, da in dieser Arbeit das Merkmal Pause vor dem Schlüsselwort als weiterer prosodischer Faktor nicht berücksichtigt wurde. Augenmerk verdient in Zukunft die Überführung des Systems von Laborbedingungen in den praktischen Einsatz und die Untersuchung von Methoden der Sprecheradaption für eine verfeinerte prosodische Modellierung.

# Literatur


[1] Bou-Ghazale S. und Asadi A.: Hands-free Voice Activation of Personal Communication Devices. In: Proc. ICASSP, Istanbul 2000

[2] Iso-Sipilä, J. und Laurila, K. und Hariharan, R. und Viiki, O.: Hands-free Voice Activation in Noisy Car Environment. In: Proc. Eurospeech, Budapest 1999.

[3] Yamashita, Y. und Mizoguchi, R.: Keyword Spotting Using F0 Contour Matching. In: Proc. Eurospeech, Rhodes 1997.

[4] Ljolje, A. und Fallside, F.: Recognition of isolated prosodic patterns using Hidden Markov Models. In: Computer, Speech and Language, pp. 27 - 33, 1987.

[5] Eichner, M. und Wolff, M. und Hoffmann, R.: A Unified Approach for Speech Synthesis and Speech Recognition Using Stochastic Markov Graphs. In: Proc. ICSLP, Beijing, 2000.

[6] Wahlster, W. editor Verbmobil: Foundations of Speech-to-Speech Translation. Springer-Verlag Berlin-Heidelberg, 2000.

[7] Wolfertstetter, F. und Ruske, G.: Structured Markov models for speech recognition. In: Proc. ICASSP, pp. 544 - 547, 1995.

[8] Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. In: Proc. IEEE, Vol. 77, pp. 257 - 286, 1989.

[9] Eichner, M. und Kühne, M. und Werner, S. und Wolff, M.: Sprachtechnologien in der Lernumgebung eines Internet-basierten Studienganges. In: Proc .ESSV, pp. 370–377, Karlsruhe, 2003.

[10] Kühne, M.: Stimmaktivierung eines Spracherkenners. Diplomarbeit. Insititut für Akustik und Sprachkommunikation, Technische Universität Dresden, 2004.