Experimente zur automatischen Schätzung der Sprechgeschwindigkeit für Spontansprache

Studienarbeit

von

Petra Philips

3. Juli 1997

Institut für Logik, Komplexität und Deduktionssysteme
Fakultät für Informatik
Universität Karlsruhe (TH)
D-76128 Karlsruhe

Betreuer:

Prof. Alexander Waibel
Dipl. Phys. Thomas Kemp

Zusammenfassung:

Bei schnell gesprochenen Äußerungen nimmt die Wortfehlerrate von Spracherkennern deutlich zu. Gelingt daher die Schätzung der Sprechgeschwindigkeit vor der eigentlichen Spracherkennung, so kann mit dieser Zusatzinformation die Erkennungsleistung verbessert werden. In dieser Arbeit wurden hauptsächlich zwei Ansätze untersucht, Sprechgeschwindigkeit für spontan gesprochene Sprache zu schätzen: durch Segmentieren mit Hilfe der Hypothese von Spracherkennern und durch Verfolgen des Entropieverlaufs der akustischen Emissionswahrscheinlichkeiten von HMM-Zuständen eines kontextunabhängigen Phonemerkenners.

Mit der Hypothese eines leistungsfähigen Spracherkenners ergibt sich eine relativ zuverlässige, wenn auch zeitaufwendige, Schätzung. Der Entropieverlauf läßt sich wesentlich schneller bestimmen. Experimentell wurde jedoch gezeigt, daß der Verlauf zu instabil ist, um eine gute Schätzung der Sprechgeschwindigkeit zu ermöglichen.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

Immer noch ist die Erkennungsleistung von *state-of-the-art*-Spracherkennern weit unter der menschlichen Spracherkennungsfähigkeit. Es ist bekannt, daß die Erkennungsleistung automatischer Spracherkenner bei stark von der Norm abweichender Sprechgeschwindigkeit deutlich abnimmt ([9, 17]). Dies hat hauptsachlich drei Ursachen: unterschiedliche Akustik der Phoneme aufgrund höherer Koartikulation, nichtpassende Wortmodelle und eine schlechte Längenmodellierung bei zu kurz gesprochenen Lauten ([9]).

Deshalb erscheint es sinnvoll, eine Schätzung der Sprechgeschwindigkeit vor der Erkennung vorzunehmen, um die Erkennungsleistung zu verbessern. Zum Beispiel kann man mehrere Spracherkenner parallel verwenden, die unterschiedlich trainiert oder mit unterschiedlichem Vokabular ausgestattet sind. Die Sprechgeschwindigkeit kann dann als Vorinformation dienen, den am besten geeigneten davon auszuwählen. Sie kann außerdem auch während der Erkennung als zusätzliches Kriterium zur Entscheidung zwischen den besten Hypothesen eines Spracherkenners eingesetzt werden ([13]). Möglich ist auch die Anpassung der Frame-Größe in der Vorverarbeitung an die Sprechgeschwindigkeit.

In den Arbeiten von Mirghafori und Siegler wird versucht, die Sprechgeschwindigkeit anhand der Hypothesen eines Spracherkenners, mit denen ein *forced alignment* durchgeführt wird, zu berechnen ([9, 17]). Die Messungen wurden auf Daten aus der WSJ-Datenbank durchgeführt. Diese enthält vorgelesene Texte und deshalb nur kleine Sprechgeschwindigkeitsschwankungen im Vergleich zu Spontansprache. Aus diesem Grund wurden in der vorliegenden Arbeit die Messungen zum Vergleich auf GSST-Daten, also spontanen Dialogen zur Terminabsprache, wiederholt. Weiter wurden auch andere Berechnungsarten für die Sprechgeschwindigkeit untersucht. Dabei stand im Vordergrund, daß man einen Zusammenhang zwischen einer großen auf diese Art ermittelten Sprechgeschwindigkeit und einer erhöhten Fehlerrate eines Spracherkenners messen kann.

Da die Schätzung aufgrund der Hypothese eines Spracherkenners zeitaufwendig und gerade bei schnellen Äußerungen häufig falsch ist, wurde versucht, mit einer neuen Methode Sprechgeschwindigkeit zu schätzen. Experimentell untersucht wurde, inwiefern der Entropieverlauf der akustischen Emissionswahrscheinlichkeiten eines Spracherkenners Hinweise auf die Sprechgeschwindigkeit liefert. Für einen kontextunabhängigen Phonemerkenner ist die akustische Entropie pro Frame ein Maß für die Unsicherheit des Erkenners darüber, welches der möglichen akustischen Ereignisse eingetreten ist (d.h. für die Unsicherheit bei der Auswahl des Phonems als Hypothese für dieses Frame). Deshalb wurde überprüft, ob bei einem geeignet trainierten Phonemerkenner die Maxima der Entropie den Phonemgrenzen entsprechen (die Unsicherheit ist groß, da ja gerade ein Übergang stattfindet).

Die Arbeit ist folgendermaßen gegliedert:
Im Kapitel 2 werden unterschiedliche Möglichkeiten vorgestellt, um Sprechgeschwindigkeit als Vorinformation für einen Spracherkenner zu berechnen. Danach werden die verwendeten Spracherkenner und die Trainings- und Testmenge in Kapitel 3 genau aufgeführt. Kapitel 4 beschreibt, wie die tatsächliche Sprechgeschwindigeit bestimmt wird. Die Experimente zur Schätzung der Sprechgeschwindigkeit sind in Kapitel 5 enthalten. Schließlich wird die Arbeit in Kapitel 6 zusammengefaßt und auf Möglichkeiten hingewiesen, die Untersuchungen weiterzuführen.

2 Sprechgeschwindigkeit

Die Sprechgeschwindigkeit (*Speech Rate, Speaking Rate, SR*) hängt hauptsächlich von zwei Faktoren ab: dem Sprecher — dessen individuellem Sprechstil (bedächtig oder hastig) und dessen emotionalen Zustand — und von der Art und dem Inhalt des Gesprochenen. Laut Siegler ist der Mittelwert der Sprechgeschwindigkeit eine Funktion des Sprechers ([17]). Die Varianz der Sprechgeschwindigkeit ist eine Funktion der kognitiven Schwierigkeit (des benötigten Denkaufwandes) des Textes. Bei einem gelesenem Text mit geringer (kognitiver) Schwierigkeit ist die Sprechgeschwindigkeit nahezu konstant. Bei Spontansprache dagegen können auch innerhalb des gleichen Satzes große Unterschiede auftreten.

Es gibt unterschiedliche Möglichkeiten die Sprechgeschwindigkeit zu messen. Zum Beispiel kann die Sprechgeschwindigkeit als die Anzahl der Worte, Silben, Vokale oder Phoneme pro Sekunde berechnen werden. Diese Maße heißen absolute Maße. Relative Maße sind zum Beispiel die gemessene Länge eines Phonems im Verhältnis zur mittleren zu erwartenden Länge dieses Phonems (in gegebenem phonetischen Kontext). Man kann die Sprechgeschwindigkeit für einen Sprecher, für einen Satz, oder für ein gegebenes Zeitintervall berechnen.

Damit die Sprechgeschwindigkeit eine sinnvolle Information für einen Spracherkenner darstellt, sollte ein Maß gewählt werden, das in nachweisbarem Zusammenhang mit der Fehlerrate steht, da diese ja reduziert werden soll. Außerdem soll sie schnell vor der eigentlichen Erkennung bestimmbar sein und einerseits zwischen schnellen und langsamen Sprechern unterscheiden, andererseits die Geschwindigkeitsunterschiede innerhalb eines Satzes deutlich machen.

2.1 Fenstergröße

Man kann die Sprechgeschwindigkeit über

- ganze Sätze
- Zeitfenster variabler Größe (getrennt z.B. durch Stille-Regionen)

  oder
- Zeitfenster konstanter Größe (überlappend)

als Referenzeinheit berechnen.

In den Arbeiten von Mirghafori und Siegler wird Sprechgeschwindigkeit für jeweils (kurze) Sätze ermittelt ([9, 17]). Dies ist für den praxisnahen Fall, in dem keine Satzgrenzen von vonherein bekannt sind, ungeeignet. Für die Berechnung der Sprechgeschwindigkeit für jeweils unterschiedlich große zusammenhängende Sprachbereiche,

getrennt durch Stille-Teile, ist eine korrekte Stilledetektion notwendig. Bei Dialogen, falls keine Sprechergrenzen bekannt sind, ergeben sich unter Umständen sehr große Fenster mit Sprache von mehreren, unterschiedlich schnell sprechenden Personen, die sich ohne Pausen abwechseln. Die allgemeinste und einfachste Methode ist deshalb die Wahl von gleich großen, sich überlappenden Zeitfenstern. Dabei werden keinerlei Vorinformation wie Transkription, Satz- und Sprechergrenzen benötigt. Allerdings besteht dabei der Nachteil, daß Phoneme durch die willkürlich gewählten Fenstergrenzen geteilt werden.

2.2 Maßeinheiten

Absolute Sprechgeschwindigkeit

Oftmals wird Sprechgeschwindigkeit in

- Anzahl Worte pro Sekunde
- Anzahl Silben pro Sekunde
- Anzahl Phoneme pro Sekunde

  oder
- Anzahl Vokale pro Sekunde

berechnet.

Man will Sprechgeschwindigkeit für relativ kurze Zeitfenster schätzen, ohne zu wissen, was genau gesprochen wurde, also *vor* der vollständigen Erkennung. Da in spontaner Sprache meistens keine Pausen zwischen Worten gemacht werden, ist es nur dann möglich, die Worte zu zählen, falls sie schon erkannt wurden. Eine wortbasierte Sprechgeschwindigkeit kann also nicht auf dem akustischen Signal vor der eigentlichen Erkennung, sondern nur danach mit Hinzunahme des Vokabulars geschätzt werden. Zusätzlich ist die wortbasierte Sprechgeschwindigkeit für kurze Sätze kaum mit der Artikulationsgeschwindigkeit korreliert und stark von den Worten abhängig („Recognize speech“ vs. „Wreck a nice beach“) ([9]).

In der Arbeit von Siegler wird gemessen, daß die Fehlerrate eines HMM-basierten Erkenners deutlich stärker mit Variationen der phonembasierten Sprechgeschwindigkeit korreliert ist als mit denen der wortbasierten ([17]). Phonemgrenzen sind unter Umständen aufgrund der Veränderung der akustischen Merkmale identifizierbar oder mit Hilfe eines kompakten, schnellen Phonemerkenners, der zeitlich nicht stark ins Gewicht fällt, detektierbar. Darauf aufbauend wird in der vorliegenden Arbeit die phonembasierte Sprechgeschwindigkeit untersucht, zumal Phoneme die akustischen Basiseinheiten sowohl für Menschen als auch für die meisten Spracherkenner sind. Da

unterschiedliche Phoneme unterschiedliche spezifische Längen haben (typisch zwischen 50 und 130 ms), muß das gewählte Zeitfenster genügend groß sein, damit die Klassifikation, ob schnell oder langsam gesprochen wurde, davon unabhängig wird.

Es scheint, daß auch die Silbengeschwindigkeit stark mit der Artikulationsrate im Deutschen und Englischen und mit der Fehlerrate von Spracherkennern zusammenhängt ([19]). Eine Silbenkerndetektion kann auf dem akustischen Signal durch Unterscheidung von stimmhaften und stimmlosen Regionen oder durch Identifizieren der zentralen Vokale erfolgen. Deshalb werden auch die beiden Maßeinheiten Vokale pro Sekunde und stimmhafte Phoneme pro Sekunde in dieser Arbeit untersucht.

Relative Sprechgeschwindigkeit

Es sind auch relative Maße möglich, wie

- gemessene Länge eines Wortes / mittlere zu erwartende Länge dieses Wortes
- gemessene Länge eines Vokals / mittlere zu erwartende Länge dieses Vokals (in einem bestimmten phonetischen Kontext)

  oder

- gemessene Länge eines Phonems / mittlere zu erwartende Länge dieses Phonems (in einem bestimmten phonetischen Kontext).

Schaaf und Kemp bzw. Anastakos verwenden relative wort- bzw. phonembasierte Maße als Vertrauensmaß nach der Erkennung, Siegler phonembasierte als Vorinformation für einen Erkenner ([13, 1, 17]). Beide weisen eine deutliche Korrelation mit der Fehlerrate bei der Erkennung auf ([13, 17]).

Der Vorteil der relativen Maße im Vergleich zu den absoluten ist der, daß dabei von den unterschiedlichen Wortlängen und unterschiedlichen spezifischen Phonemlängen (typisch zwischen 50 und 130 ms - abhängig auch vom Kontext) abstrahiert wird. Allerdings müssen die mittleren zu erwartenden Längen im voraus mit einem höheren Aufwand anhand von Trainingsdaten geschätzt werden.

Wie die absolute wortbasierte Geschwindigkeit hängt die relative von der Korrektheit der erkannten Worte ab, ist also vor der eigentlichen Erkennung nicht schätzbar. Relative kontextabhängige phonembasierte Maße benötigen einen höheren Aufwand für das Schätzen der mittleren erwarteten Längen und sind fehleranfällig, falls der Kontext (linkes und rechtes Phonem) falsch erkannt wird.

Siegler verweist darauf, daß Vokallängen die stärksten Schwankungen abhängig von der Sprechgeschwindigkeit aufweisen ([17]). Es ist deshalb interessant zu überprüfen, ob relative Vokallängen eine gute und genau zu schätzende Sprechgeschwindigkeit sind.

2.3 Pausen und Geräusche

Um den Zusammenhang zwischen der Sprechgeschwindigkeit und deren Schätzung korrekt nachzuweisen, werden in dieser Arbeit wie in [9, 17, 18] Pausen und nichtsprachliche Geräusche innerhalb der Äußerungen weggeschnitten.

Fazit

Im realitätsnahen Fall, in dem keinerlei Vorinformation über die akustische Eingabe bekannt ist, ist es am einfachsten, Sprechgeschwindigkeit für Zeitfenster einer konstanten Größe zu approximieren. Wortbasierte Maße benötigen die (korrekt) erkannten Worte, sind also vor einer Erkennung nicht direkt auf dem akustischen Signal schätzbar. Phonem- und Silben(kern)basierte Maße entsprechen eher der Artikulationsgeschwindigkeit und sind stärker mit der Fehlerrate von Spracherkennern korreliert. Absolute Maße sind mit einem kleineren Aufwand zu schätzen und benötigen weniger Vorinformation. Um der reinen Artikulationsrate nahezukommen, ist es sinnvoll, eine schnelle Pausendetektion vor der Schätzung durchzuführen, und nur auf Sprachdaten zu schätzen. Zu beachten ist auf jeden Fall, daß die berechneten Werte für die Sprechgeschwindigkeit von den gewählten Phonem- und Silbenmodellen und von dem Berechnungsmodus abhängen.

3 Spracherkenner und Datenbasis

Für die in Kapitel 5 beschriebenen Experimente wurde das Janus-3 Spracherkennungssystem eingesetzt. Für Merkmalsextraktion, akustische Modellierung, Sprachmodellierung und Suche existieren konfigurierbare und kombinierbare Module. Aufgrund der Einbettung in Tcl/Tk, in der die benötigten Funktionen in einer objektorientierten Weise als Benutzerschnittstelle zugänglich sind, ist es möglich, eigene Erweiterungen zu integrieren.

3.1 Kontextabhängiger Worterkenner

Ein Teil der Messungen wurden mit dem kontextabhängigen System durchgeführt, welches bei der Verbmobil-Evaluation 1996 auf der GSST-Datenbasis eine Wortfehlerrate von 13,8% erzielte und in [2] ausführlich beschrieben wird.

Die Eingabe ist ein mit 16kHz abgetastetes 16 Bit-kodiertes Sprachsignal (*high quality close-speaking microphone*). Die Frame-Rate bei der Signalvorverarbeitung beträgt 100Hz. Für jedes 16ms-Zeitfenster wird aufgrund einer 256-Punkt Fouriertransformation das Leistungsspektrum berechnet. Die 129 erhaltenen Parameter werden nach einer Vokaltraktlängennormalisierung (([8, 20]) zu 30 mel-Koeffizienten zusammengefaßt, logarithmiert, und mittels Kosinustransformation zu 13 mel-Cepstrum-Parametern umgerechnet ([14]). Danach erfolgt eine Kanalnormalisierung durch Abziehen des Mittelwertes des mel-Cepstrums des gesamten Sprachteils (ohne Stille) der Äußerung. Der 39-dimensionale Merkmalsvektor, der aus den aneinandergereihten normalisierten mel-Cepstrum-Parametern entsteht, sowie deren erste und zweite Differenzen, wird mittels einer linearen Diskriminanzanalyse (LDA) transformiert und auf 32 Dimensionen reduziert ([14]).

Die akustischen Einheiten sind Pentaphone, sie werden mit Hidden Markov Modellen durch je 6 HMM-Zustände, wie in Abbildung, 1 modelliert (drei unterschiedliche Zustände für Anfang(b), Mitte(m) und Ende(e)). Das Stille-Modell besteht aus vier identischen Zuständen.

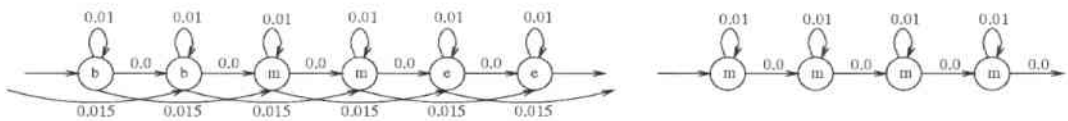


Abbildung 1: Topologie des HMMs für Phoneme und Stille (angegebene Übergangswahrscheinlichkeiten sind logarithmiert und negiert)

Die Emmisionswahrscheinlichkeitsdichten der HMM-Zustände werden als multivariate Gaußverteilungen, jede eine Linearkombination von 16 (32-dimensionalen) Normalverteilungen mit diagonalen Varianzen, geschätzt. Insgesamt gibt es 10.018 akustische Modelle, die sich 2518 Mischverteilungen teilen, welche nach einer zweistufi-

gen Ballungsanalyse ([4]) erhalten wurden. Eine genaue Liste der modellierten Phoneme findet sich in Tabelle 3.

Bei der Baumsuche werden unterschiedliche Sprachmodelle eingesetzt: zu der stochastischen Trigramm-Grammatik (linear interpoliert mit klassenbasierten Modellen) werden ein zusätzliches Buchstabiermodell und ein Bigrammmodell, das auf linguistische Kategorien basiert, hinzugezogen. Der Wortschatz enthält 5748 Worte, inklusive Geräusche und einzelne Buchstaben.

Das System wurde 5 Iterationen auf GSST-Daten der Gesamtdauer von über 33 Stunden trainiert (siehe Tabelle 1).

3.2 Kontextabhängiger Phonemerkenner

Der verwendete kontextabhängige Phonemerkenner unterscheidet sich von dem kontextabhängigen Worterkenner darin, daß sein Wortschatz nur aus Phonemen besteht, und das Sprachmodell jedes Phonem zu jedem Zeitpunkt als gleich wahrscheinlich zuläßt. Die Erkennung erfolgt also nur aufgrund des akustischen Modells.

3.3 Kontextunabhängiger Phonemerkenner

Die Vorverarbeitung ist die gleiche wie bei den anderen beiden Systemen, nur daß keine Vokaltraktlängennormalisierung vorgenommen wird. Für die akustischen Basiseinheiten des kontinuierlichen HMMs (die gleichen Kernphoneme wie oben, diesmal ohne Kontext), werden 199 Mischverteilungen (3*66 Phoneme + Stille) geschätzt, wiederum als Linerakombination von 16 Normalverteilungen. Die Topologie der HMM-Modelle für die Phoneme ist die gleiche wie in Abbildung 1. Das Modell für Stille besteht aus nur 2 identischen Zuständen. Auch hier wird das Phonemgleichverteilungs-Sprachmodell verwendet. Der Erkenner wurde 3 Iterationen auf GSST-Äußerungen der Gesamtdauer von über 30 Stunden trainiert (s. Tabelle 1).

3.4 Datenbasis

Die verwendeten Sprachdaten sind eine Teilmenge der GSST-Datenbasis (German Spontaneous Scheduling Task). Diese Datenbasis enthält Dialoge zur Terminabsprache, die mit 16kHz Abtastrate aufgenommen wurden. Die Äußerungen enthalten aufgrund ihrer Spontaneität auch grammatikalisch nicht korrekte und unvollständige Sätze, Stottern, Buchstabierungen, Pausen und verschiedene Geräusche, wie z.B. Atmen, Schmatzen, Husten, Lachen oder mechanisches Klicken.

Für das Trainieren der verschiedenen Erkenner wurden die in Tabelle 1 beschriebenen Daten verwendet.

System	Datenmenge	Sprecher	Äußerungen	Dauer(min)
Kontextabhängig	männlich	409	8318	1265
	weiblich	278	5343	739
	gesamt	687	13661	2004
Kontextunabhängig	männlich	368	7621	1173
	weiblich	247	4792	669
	gesamt	615	12413	1842

Tabelle 1: Trainingsdatenmengen für die kontextabhängigen bzw. -unabhängigen Systeme

Für die Untersuchungen wurden insgesamt 436 Äusserungen mit einer gesamten Dauer von 66 Minuten, davon 44 Minuten reine Sprache (67%) und 22 Minuten Pausen und Geräusche (33%), verwendet. Diese Daten wurden nicht zum Trainieren der Erkenner verwendet. Eine genaue Beschreibung der Testmenge ist in Tabelle 2 gegeben.

Testmenge	Sprecher	Äußerungen	Dauer(min)	Reine Sprache(min)
männlich	66	255	38	24
weiblich	39	181	28	20
gesamt	105	436	66	44

Tabelle 2: Testdatenmenge für die Schätzung der Sprechgeschwindigkeit

4 Tatsächliche Sprechgeschwindigkeit

4.1 Finden der Phonemgrenzen

Um Schätzungen für die Sprechgeschwindigkeit auf ihre Genauigkeit zu überprüfen, muß die tatsächliche Sprechgeschwindigkeit bekannt sein. Dafür sind genaue Phonemsegmentierungen der Äußerungen notwendig. Handsegmentierte Daten wären wünschenswert, sind aber im benötigten Umfang nicht vorhanden. Deshalb wurde für eine automatische Segmentierung ein *forced Viterbi alignment* mit dem kontextabhängigen Spracherkenner auf der korrekten Worttranskription durchgeführt.

Obwohl die Segmentierung meistens genau ist (bis auf 1-2 Frames), kommt es in manchen Fällen vor, daß an der Grenze zu Stille-Regionen sehr große Fehler auftreten, was die Korrelationswerte verfälscht, z.B. für die Äußerung „okay +KLICK+" (mdh1_mjk3_tsponti2_mjk3_2_08):

{ O { 0 2 } }
{ U { 3 5 } }
{ K { 6 8 } }
{ E { 9 11 } }
{ I { 12 328 } }
{ +nKL { 329 331 } }

Die Segmentierung ergibt für jedes Phonem mindestens 30 Millisekunden Länge (3 Frames). Dabei kann es sein, daß in Wirklichkeit die Länge mancher Phoneme bei sehr schnellen Sprechern kürzer ist, oder daß die Laute ganz verschluckt werden. Diese Ungenauigkeiten sind ohne handsegmentierte Daten nicht vermeidbar.

4.2 Berechnen der tatsächlichen Sprechgeschwindigkeit

Um nicht auf kurze Äußerungen eingeschränkt zu sein, wurde jeweils ein Wert für sich überlappende Zeitfenster konstanter Länge bestimmt. In der Testdatenmenge beträgt die mittlere Dauer der Worte (9286 an der Zahl) 0,29 Sekunden. Deshalb wurden Fenster der Längen:

- 0,5 Sekunden (50 Frames)
- 1 Sekunde (100 Frames)
- 2 Sekunden (200 Frames)

untersucht, die sich jeweils um die Hälfte überlappen (s. Abbildung 2).

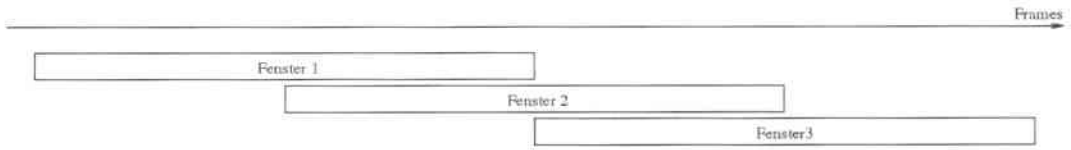


Abbildung 2: Zeitfenster konstanter Größe, die sich jeweils um die Hälfte überlappen

Nachdem Pausen und Geräusche ausgeschnitten wurden, wurde die Sprechgeschwindigkeit für jedes Fenster mit folgenden Formeln berechnet (es werden jeweils nur vollständige, nicht von den Fenstergrenzen gekürzte Phoneme gezählt):

- P = Anzahl Phoneme / Gesamtdauer Phoneme
- V = Anzahl Vokale / Gesamtdauer Vokale
- VD = Anzahl Vokale und Vokaldiphtonge / Gesamtdauer Vokale und Vokaldiphtonge

Falls die Fenstergrenzen mit den Phonemgrenzen übereinstimmen erhält man für P genau die in der Literatur ([9, 17]) untersuchte Sprechgeschwindigkeit IMD (Inverse of Mean Duration) ($= \frac{n}{\sum_{i=1}^{n} d_i}$, wobei n die Anzahl der Phoneme im Fenster und d_i die Längen der Phoneme in Sekunden sind). Im allgemeinen fallen die Phonemgrenzen natürlich nicht mit den künstlich gewählten Fenstergrenzen zusammen. Sieht man die innerhalb der Fenstergrenzen fallenden Sprachteile als eine Äußerung an, lassen sich die Ergebnisse mit denen aus den Arbeiten von Siegler ([17]) und Mirghafori ([9]) vergleichen.

Da Vokallängen die größten Schwankungen abhängig von der Sprechgeschwindigkeit aufweisen ([17]), wurde untersucht, inwiefern es einfacher ist, nur Vokale in der Schätzung zu berücksichtigen. Es wurden zwei Methoden V und VD für die Zählung der Vokale untersucht. Für V werden nur die einfachen Vokale gezählt, für VD zusätzlich Vokaldiphtonge (siehe Tabelle 3).

Für der Schätzung der Phonemübergänge aufgrund der akustischen Entropie (s. Abschnitt 5.3.1) wurden die Phonemübergänge in zwei Varianten gezählt (wobei Diphtonge einmal 1 mal und einmal 2 mal gezählt wurden):

- PZ = Anzahl Phonemanfänge/Fensterlänge
- PDZ = Anzahl Phonemanfänge + Anzahl Diphtonganfänge/Fensterlänge

Die Silbengeschwindigkeit wurde auch noch durch einfache Zählung der Vokale approximiert:

- VZ = Anzahl Vokalanfänge/Fensterlänge

- VDZ = Anzahl Vokalanfänge + Anzahl Vokaldiphtonge/Fensterlänge

Tabelle 3 faßt die in dieser Arbeit verwendete Phonemmenge sowie die Untermengen der Vokale und stimmhaften Laute zusammen. Sie wurden aus den Phonemklassen des Verbmobil-Erkenners übernommen.

Klasse	Phoneme
Phoneme	A AR AEH AEHR AH AHR AI AU B CH X D E E2 EH EHR ER ER2 EU F G H I IR I E IHR J K L M N NG O OR OE OEH ANG OH OHR P R S SCH T TS TSCH U UR UE UEH UEHR UH UHR V Z SIL +QK +hBR +hEH +hEM +hGH +hHM +hLG +hSM +nGN +nKL +nMK
Stille	SIL
Geräusche	+QK +hBR +hEH +hEM +hGH +hHM +hLG +hSM +nGN +nKL +nMK
Diphtonge	AR AEHR AHR AI AU EHR ER EU IR IHR OR OHR TS TSCH UR UEHR UHR
Vokale	A AEH AH E E2 EH ER2 I IE O OE OEH ANG OH U UE UEH UH
Vokaldiphtonge	AR AEHR AHR AI AU EHR ER EU IR IHR OR OHR UR UEHR UHR

Tabelle 3: Verwendete Phonemklassen

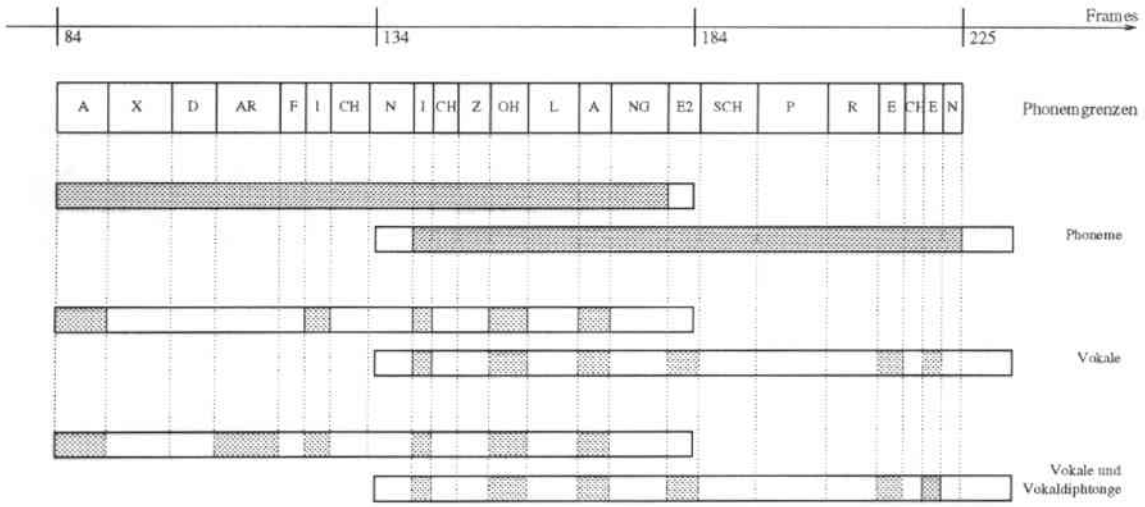
Zur Illustrierung betrachte man die Äußerung, deren Transkription lautet: „ach darf ich nich' so lange sprechen". Die Phonemsegmentierung (mit *forced Viterbi alignment* aufgrund der Transkription, die Framegrenzen in Klammern) sieht folgendermaßen aus:

{ SIL { 0 83 } }
{ A { 84 91 } }
{ X { 92 101 } }
{ D { 102 108 } }
{ AR { 109 118 } }
{ F { 119 122 } }
{ I { 123 126 } }
{ CH { 127 132 } }
{ N { 133 139 } }
{ I { 140 142 } }
{ CH { 143 146 } }
{ Z { 147 151 } }
{ OH { 152 157 } }
{ L { 158 165 } }
{ A { 166 170 } }
{ NG { 171 179 } }

{ E2 { 180 184 } }
{ SCH { 185 193 } }
{ P { 194 204 } }
{ R { 205 212 } }
{ E { 213 215 } }
{ CH { 216 218 } }
{ E2 { 219 221 } }
{ N { 222 224 } }
{ SIL { 225 228 } }

Daraus ergeben sich nach Herauschneiden der Stille, z.B für Fenster der Länge 1s (100 Frames), 2 Fenster mit Sprechgeschwindigkeitswerten wie in Abbildung 3.

Die Histogramme für die auf den Testdaten gemessenen unterschiedlichen Sprechgeschwindigkeiten sind in den Abbildungen in Anhang A dargestellt. Man kann sehen, daß sie ungefähr Normalverteilungen darstellen.



P:	$\frac{15 \cdot 100}{96} = 15,62$	$\frac{15 \cdot 100}{85} = 17,64$
V:	$\frac{5 \cdot 100}{26} = 19,23$	$\frac{6 \cdot 100}{25} = 24$
VD:	$\frac{6 \cdot 100}{36} = 16,66$	$\frac{6 \cdot 100}{25} = 24$
PZ:	16	$\frac{15 \cdot 100}{91} = 16,48$
PDZ:	17	$\frac{15 \cdot 100}{91} = 16,48$
VZ:	6	$\frac{6 \cdot 100}{91} = 6,59$
VDZ:	7	$\frac{6 \cdot 100}{91} = 6,59$

Abbildung 3: Beispiel zur Berechnung der Sprechgeschwindigkeit für Fenster der Länge 1 Sekunde (100 Frames) für die Äußerung „ach darf ich nich' so lange sprechen" (fdm1_mth1_tsponti3_fdm1_3_03)

5 Schätzung der Sprechgeschwindigkeit

5.1 Zusammenhang mit der Fehlerrate

Zuerst wurden unterschiedliche Maße untersucht, inwiefern sie bei hohen Werten in Zusammenhang mit der Fehlerkennung stehen. Da die Zeitfenster klein sind und die Zeitfenstergrenzen im Inneren der Worte liegen können, war es nicht möglich, die Wortfehlerrate in Funktion der Sprechgeschwindigkeit zu berechnen. Deshalb wurden die gesamten Worte der Testmenge in zwei Klassen geteilt: die Worte, die mit dem Verbmobil-Erkenner korrekt erkannt werden, und die Worte, die nicht korrekt erkannt werden (siehe Tabelle 4).

Wortmenge	Anzahl	Dauer insgesamt (min)	mittlere Dauer (s)
korrekt erkannte	7420	38,6	0,31
falsch erkannte	1866	7,4	0,23
gesamt	9286	46	0,29

Tabelle 4: Aufschlüsselung der Worte aus der Testmenge, die der Verbmobil-Erkenner korrekt bzw. falsch erkennt

Für jedes Wort wurde die tatsächliche Sprechgeschwindigkeit mit allen in Abschnitt 4.2 beschriebenen Maßen berechnet (Fenstergröße = Länge des Wortes). Dann wurde der Mittelwert der verschiedenen tatsächlichen Sprechgeschwindigkeiten auf den beiden Wortmengen verglichen. In Tabelle 5 sieht man, daß für falsch erkannte Worte der Mittelwert der Sprechgeschwindigkeiten höher ist.

Wortmenge	P, PZ	V	VD	VZ	VDZ
korrekt erkannte	14,7 (32,2)	14,2 (81,2)	15,7 (51,8)	4,8 (10,8)	6,1 (8,2)
falsch erkannte	17,8 (57,4)	15,8 (133,0)	18,3 (93,9)	5,7 (22,9)	7,1 (18,3)

Tabelle 5: Mittelwerte der verschiedenen Sprechgeschwindigkeiten auf den korrekt vs. falsch vom Verbmobil-Erkenner erkannten Worte (Varianz in Klammer)

5.2 Schätzung aufgrund der Hypothese von Spracherkennern

5.2.1 Schätzung der Phonemgeschwindigkeit mit den kontextabhängigen Wort- und Phonemerkennern

In den Arbeiten von Siegler und Mirghafori wurde die Phonem-Sprechgeschwindigkeit (IMD, siehe Abschnitt 4.2) aufgrund der Hypothese eines guten Erkenners, mit der dann die Segmentierung erfolgte, geschätzt ([9, 17]). Die Experimente wurden auf Äußerungen aus der Datenbasis WSJ0 und WSJ1 durchgeführt. WSJ enthält

vorgelesene Texte, bei denen die Varianz der Sprechgeschwindigkeit kleiner ist als bei Spontansprache. Außerdem enthält sie keine ungrammatikalischen Äußerungen, so daß die Erkennungsleistung darauf höher ist als auf spontaner Sprache. In den genannten Arbeiten wurde die Korrelation auf ganzen Äußerungen der mittleren Länge von 7 Sekunden gemessen, was zu einer "Verschmierung" der Segmentierungsungenauigkeiten führt. In beiden Arbeiten wurde eine hohe Korrelation (0,93) der so geschätzten Geschwindigkeit mit der tatsächlichen gemessen. Ebenso wurde eine niedrigere, aber immer noch gute lineare Korrelation (0,7) aufgrund der Hypothese eines kontextabhängigen Phonemerkenners gemessen.

In der vorliegenden Arbeit wurden die Messungen zum Vergleich für Spontansprache für das Maß P (Anzahl Phoneme / Gesamtdauer Phoneme) wiederholt.

Das Problem bei der Schätzung aufgrund der Hypothese eines Erkenners ist, daß gerade schnelle Äußerungen zu hohen Fehlerraten führen. Dadurch wird wiederum die Schätzung der Sprechgeschwindigkeit unzuverlässig. Um zu überprüfen, inwiefern gerade die schnelleren Äußerungen zuverlässig geschätzt wurden, wurde zusätzlich die Korrelation für eine Teilmenge aus den schnellsten Fenstern ermittelt. Es wurden diejenigen Fenster gesondert betrachtet, für die der tatsächliche Sprechgeschwindigkeitswert größer als der Mittelwert + 1,5 mal Standardabweichung ist (ungefähr 5-8 % der Werte). Schon bei um einer Standardabweichung vom Mittelwert liegende Werte nimmt die Fehlerrate eines kontextabhängigen Erkenners laut Siegler ([17]) deutlich zu.

Die gemessenen Korrelationskoeffizienten zwischen den geschätzten Sprechgeschwindigkeiten und den tatsächlichen auf GSST-Daten sind in Tabelle 6 aufgeführt. In Abbildung 4 sind die dazugehörenden Scatterplots gezeichnet. Auf der Abszisse wurde die tatsächliche, auf der Ordinate die dazu gehörende geschätzte Sprechgeschwindigkeit aufgetragen. Der waagerechte und der senkrechte Strich zeigen die Klasssengrenze langsam/schnell an. Die Regressionsgerade für die gesamte Testmenge ist gestrichelt gezeichnet.

System	Testmenge	Fenster 0,5s	Fenster 1s	Fenster 2s	[9]	[17]
Kontextabh. Worterkenner	gesamt	0,80	0,83	0,87	0,93	0,93
	schnell	0,05	0,31	-0,06		
Kontextabh. Phonemerkenner	gesamt	0,40	0,46	0,47	0,73	
	schnell	-0,05	0,30	0,56		

Tabelle 6: Korrelationskoeffizienten für Schätzung von P (Anzahl Phoneme / Gesamtdauer) aufgrund der Hypothese des kontextabhängigen Worterkenners und des Phonemerkenners. Die schnellsten 5-8 % der Äußerungen sind die, die um 0,5 mal Standardabweichung größer als der Mittelwert (tatsächliche Sprechgeschwindigkeit) sind. Die Vergleichswerte aus der Literatur sind auf WSJ gemessen, pro Äußerung, mit Fenstern von ungefähr 7 Sekunden Länge.

Die erhaltenen Korrelationskoeffizientwerte für den Worterkenner sind nur wenig niedriger als die in der Literatur vermerkten von 0.93. Die lineare Korrelation wächst

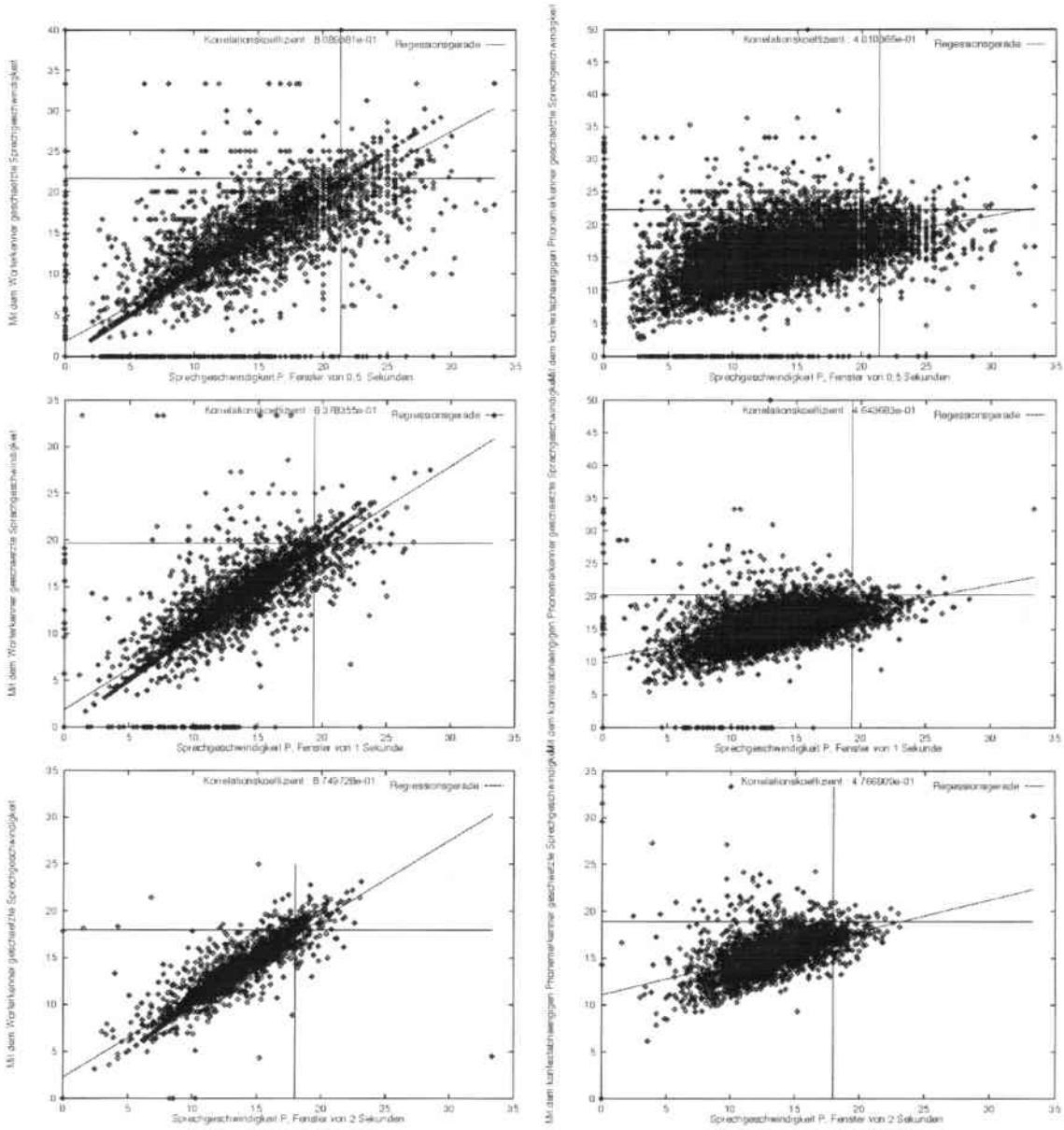


Abbildung 4: Scatterplots für die Schätzung von P (Anzahl Phoneme /Gesamtdauer) aufgrund der Hypothese des Worterkenners (linke Spalte) und des kontextabhängigen Phonemerkenners (rechte Spalte) für Fenster von 0,5, 1 und 2 Sekunden. Auf der Abszisse ist die tatsächliche, auf der Ordinate die geschätzte Sprechgeschwindigkeit aufgetragen.

mit der Fenstergröße, da dabei eine stärkere Mittelung der Sprechgeschwindigkeit erfolgt, und kleine Ungenauigkeiten und Variationen weniger Einfluß auf das Ergebnis haben. Es ist durchaus möglich, daß bei Fenstergrößen von 7 Sekunden die Korrelationswerte noch höher werden. Dabei verliert man jedoch die Variationen der Sprechgeschwindigkeit innerhalb eines Satzes. Anastakos und Siegler nennen aber gerade hohe Fluktuationen der Sprechgeschwindigkeit als eine Ursache für Erkennungsfehler ([1, 17]). Deshalb ist es sinnvoll, kleinere Fenster zu betrachten.

Für die schnellen Äußerungen ist der Korrelationskoeffizient deutlich niedriger. Man bemerkt, daß Fenster von 0,5 Sekunden Länge zu klein sind, um eine zuverlässige Schätzung zu erlauben. Ungenauigkeiten bei der Segmentierung aufgrund der Hypothese (die bei schnellen Äußerungen öfters vorkommen) haben eine große Auswirkung auf die Schätzung der Sprechgeschwindigkeit. Für Fenster der Größe von 2 Sekunden ist kein linearer Zusammenhang erkennbar, allerdings ist die Datenmenge kleiner (165 Werte) und die Signifikanz niedriger (40% Wahrscheinlichkeit, daß das Ergebnis zufällig ist ([3])). Falsche Segmentierungen (Ausreißer) haben ein stärkeres Gewicht. Für Fenster der Länge einer Sekunde (334 Werte) ist der Korrelationskoeffizient 0,3 mit einer Signifikanz von 0,99 korrekt (1% Wahrscheinlichkeit, daß das Ergebnis zufällig ist).

Für den Phonemerkenner wurden wesentlich schlechtere Ergebnisse, als in der Literatur vermerkt, erhalten. Wiederum steigt der lineare Zusammenhang mit der Fenstergröße, und die Fenstergröße von 0,5 Sekunden erweist sich als zu klein. Die besseren Korrelationswerte für das Fenster von 2 Sekunden sind ebenfalls nicht sehr signifikant, es ist allerdings aus den Scatterplots eine hohe lineare Korrelation ersichtlich.

System	Testmenge	Fenster 0,5s	Fenster 1s	Fenster 2s
Kontextabhängiger Worterkenner	gesamt	96%	95%	95%
	schnell	59%	57%	67%
	langsam	98%	98%	97%
Kontextabhängiger Phonemerkenner	gesamt	91%	91%	90%
	schnell	8%	7%	12%
	langsam	96%	96%	95%

Tabelle 7: Klassifikationsrate schnell vs. langsam aufgrund der Hypothese des kontextabhängige Worterkenners und des Phonemerkenners. Sprechgeschwindigkeit ist P (Anzahl Phoneme / Gesamtdauer der Phoneme). Schnelle Werte sind größer als Mittelwert + 1,5 mal Standardabweichung (5-6 %). Es werden die prozentual korrekt klassifizierten Werte angegeben.

Möchte man nur eine Klassifikation nach schnell oder langsam, um danach eventuell auf neue akustische Modelle oder Wörterbücher bei der eigentlichen Erkennung umzusteigen, ist nicht der lineare Korrelationkoeffizient interessant, sondern die Klassifikationsrate. Wiederum wurde die Grenze schnell/langsam bei Mittelwert + 1.5 Standardabweichung gesetzt (tatsächliche Sprechgeschwindigkeit). Die Klassifikationsraten sind in Tabelle 7 aufgeführt. In den Scatterplots (Abbildung 4) enthält der rechte obere Quadrant die korrekt schnell klassifizierten, der linke untere die korrekt langsam klassifizierten Werte. Man bemerkt, daß die Ergebnisse für den Worterkenner akzeptabel sind. Für Online-Erkennung ist ein gesamter zusätzlicher Erkennungslauf zu zeitintensiv (30 - 80 mal Echtzeit). Der Phonemerkenner, obwohl schneller, eignet sich allerdings kaum zur Detektierung von schnell gesprochenen Äußerungen. Die Klassifikationsrate für die schnellen Äußerungen liegt kaum über der Baseline (5-8 % der Äußerungen sind die schnellen). Bei schneller Sprache wird

die minimale Länge im HMM-Modell bei der Aussprache zum Teil unterschritten ([9]). Die Transkription kennzeichnet nicht immer „verschluckte" Laute als solche. Bei der Segmentierung mit der korrekten Transkription werden für solche Phoneme die minimale Dauer von 3 Frames gelabelt. Mit dem Worterkenner werden durch die Wörterbücher auch die fast fehlenden Phoneme erzwungen. Beim Phonemerkenner sind aber diese Laute unwahrscheinlicher, es werden weniger Phoneme von längerer Dauer bevorzugt. Schnelle tatsächliche Sprechgeschwindigkeiten werden dadurch nicht gefunden.

In der vorliegenden Arbeit wurde deshalb untersucht, inwiefern andere Maße genauer und trotzdem schneller als mit Hilfe eines Worterkenners zu schätzen seien.

5.2.2 Schätzung der Sprechgeschwindigkeit aufgrund der Hypothese des kontextunabhängigen Phonemerkenners

Auf den falsch erkannten Worten wurde die tatsächliche Sprechgeschwindigkeit mit der aufgrund des kontextunabhängigen Phonemerkenners geschätzten verglichen. Die Korrelationskoeffizienten sind in Tabelle 8 angegeben. Man bemerkt, daß P am wenigsten robust ist gegenüber der falschen Erkennung. Vokale werden zuverlässiger erkannt, da sie bei Fehlerkennung meistens untereinander verwechselt werden. Für PZ (die Anzahl der Phonemanfänge in einem Fenster) ergibt sich eine höhere lineare Korrelation. Dies liegt daran, daß durch einfache Zählung ohne mit der Gesamtlänge zu normieren eine gröbere Quantisierung erfolgt, was zu einer stärkeren linearen Korrelation führt (z.B sind (1,1,1) (1.1,1,0.9) schwächer linear korreliert als (1,1,1) und (1,1,1)).

Die Korrelationen sind für alle Maße schwach, so daß keine besonders gute Schätzung zu erwarten ist. Die Ergebnisse sind in den Tabellen 9, 10 und 11 zu finden. In den Abbildungen 5, 6 und 7 sind die entsprechenden Scatterplots aufgetragen.

Wortmenge	P	V	VD	PZ	VZ	VDZ
falsch erkannte	0,09	0,15	0,13	0,28	0,21	0,31

Tabelle 8: Korrelationskoeffizienten zwischen der tatsächlichen und mit dem kontextunabhängigen Phonemerkenner geschätzten Sprechgeschwindigkeit auf den vom Verbmobil-Erkenner falsch erkannten Worten

Maß	Fenster 0,5s	Fenster 1s	Fenster 2s
P	0,40	0,47	0,49
V	0,31	0,33	0,41
VD	0,40	0,45	0,50
PZ	0,43	0,41	0,39
VZ	0,43	0,41	0,37
VDZ	0,49	0,48	0,45

Tabelle 9: Korrelationskoeffizienten für die Schätzung aufgrund der Hypothese des kontextunabhängigen Phonemerkenners auf der gesamten Testmenge

Maß	Fenster 0,5s	Fenster 1s	Fenster 2s
P	0,14	0,30	0,41
V	-0,11	-0,17	-0,03
VD	0,00	0,09	0,28
PZ	0,07	0,15	0,22
VZ	0,12	0,07	0,13
VDZ	0,17	0,07	0,32

Tabelle 10: Korrelationskoeffizienten für die Schätzung aufgrund der Hypothese des kontextunabhängigen Phonemerkenners auf den schnellen Fenstern

Maß	Klassifikationsrate	Fenster 0,5s	Fenster 1s	Fenster 2s
P	insgesamt korrekt	91%	90%	89%
	schnelle korrekt	8%	14%	14%
	langsame korrekt	96%	96%	94%
V	insgesamt korrekt	90%	89%	89%
	schnelle korrekt	13%	11%	16%
	langsame korrekt	94%	94%	94%
VD	insgesamt korrekt	89%	90%	89%
	schnelle korrekt	14%	15%	23%
	langsame korrekt	95%	95%	94%
PZ	insgesamt korrekt	92%	89%	90%
	schnelle korrekt	14%	11%	8%
	langsame korrekt	96%	96%	96%
VZ	insgesamt korrekt	89%	89%	87%
	schnelle korrekt	14%	18%	16%
	langsame korrekt	96%	94%	95%
VDZ	insgesamt korrekt	93%	89%	92%
	schnelle korrekt	18%	18%	24%
	langsame korrekt	95%	94%	95%

Tabelle 11: Klassifikationsrate des Klassifikators aufgrund der Hypothese des kontextunabhängigen Phonemerkenners. Es werden prozentual die korrekt klassifizierten Werte angegeben.

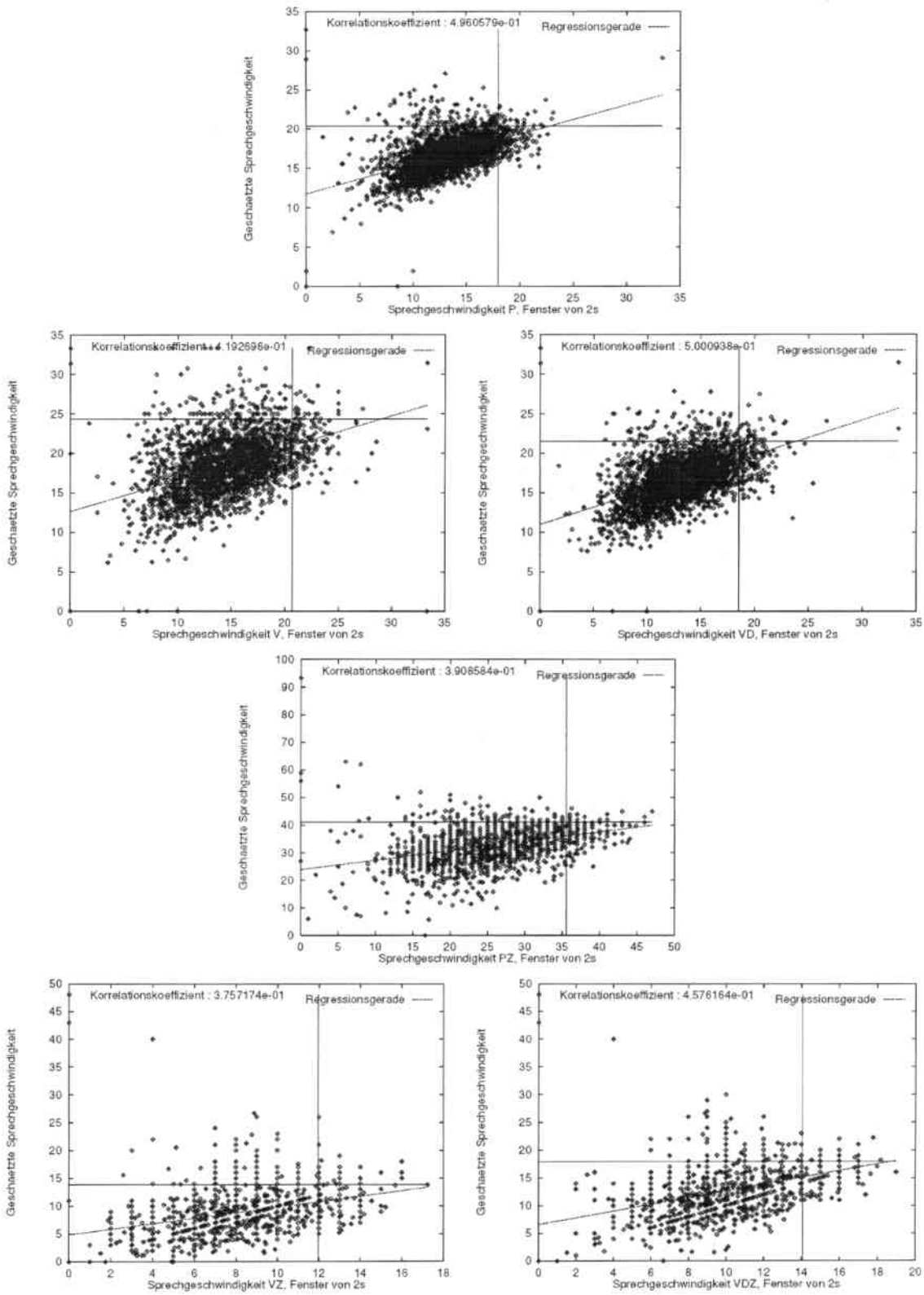


Abbildung 5: Scatterplots für die Schätzung von Sprechgeschwindigkeiten aufgrund der Hypothese des kontextunabhängigen Phonemerkenners auf Fenster von 2 Sekunden

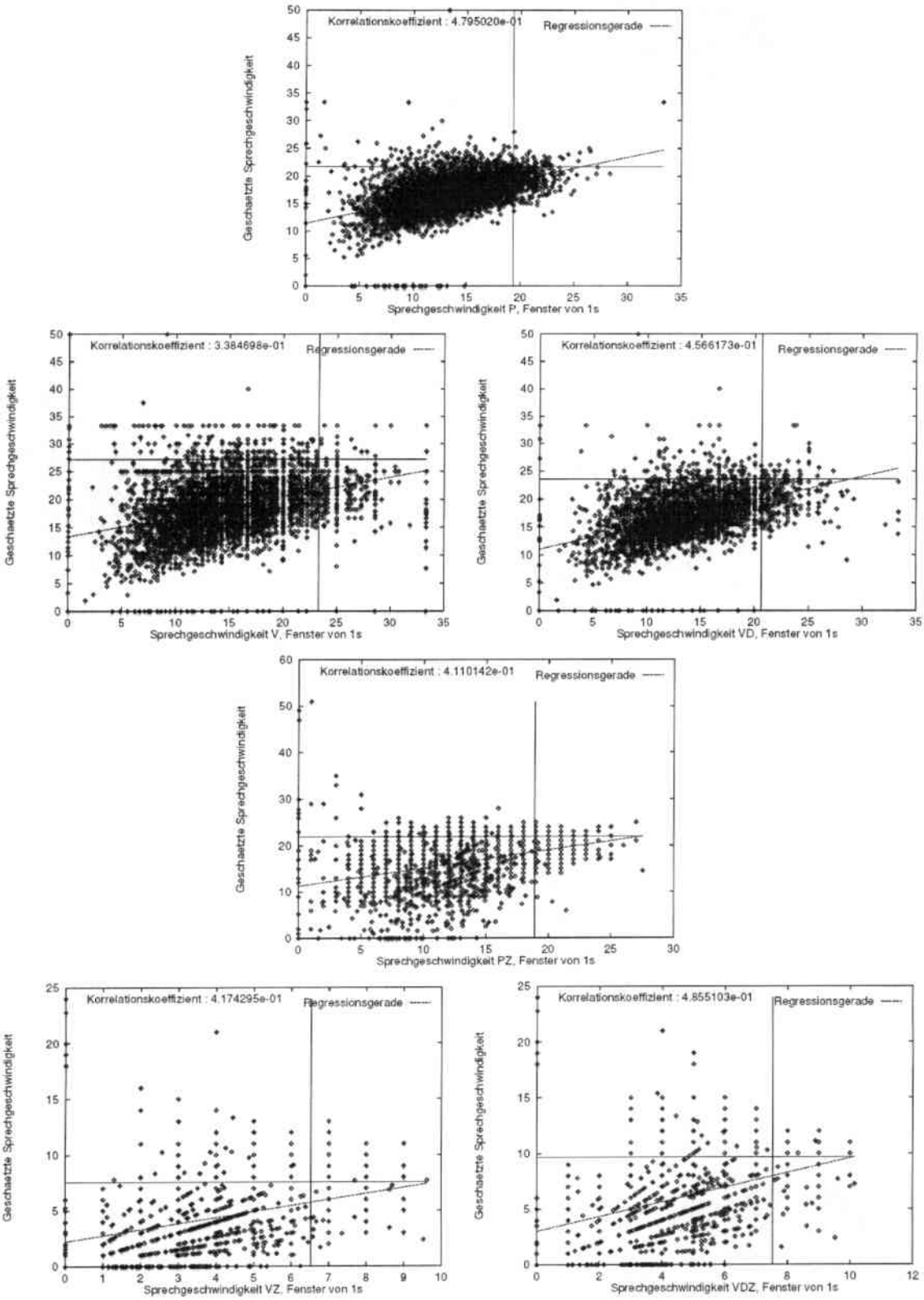


Abbildung 6: Scatterplots für die Schätzung von Sprechgeschwindigkeiten aufgrund der Hypothese des kontextunabhängigen Phonemerkenners auf Fenster von 1 Sekunde

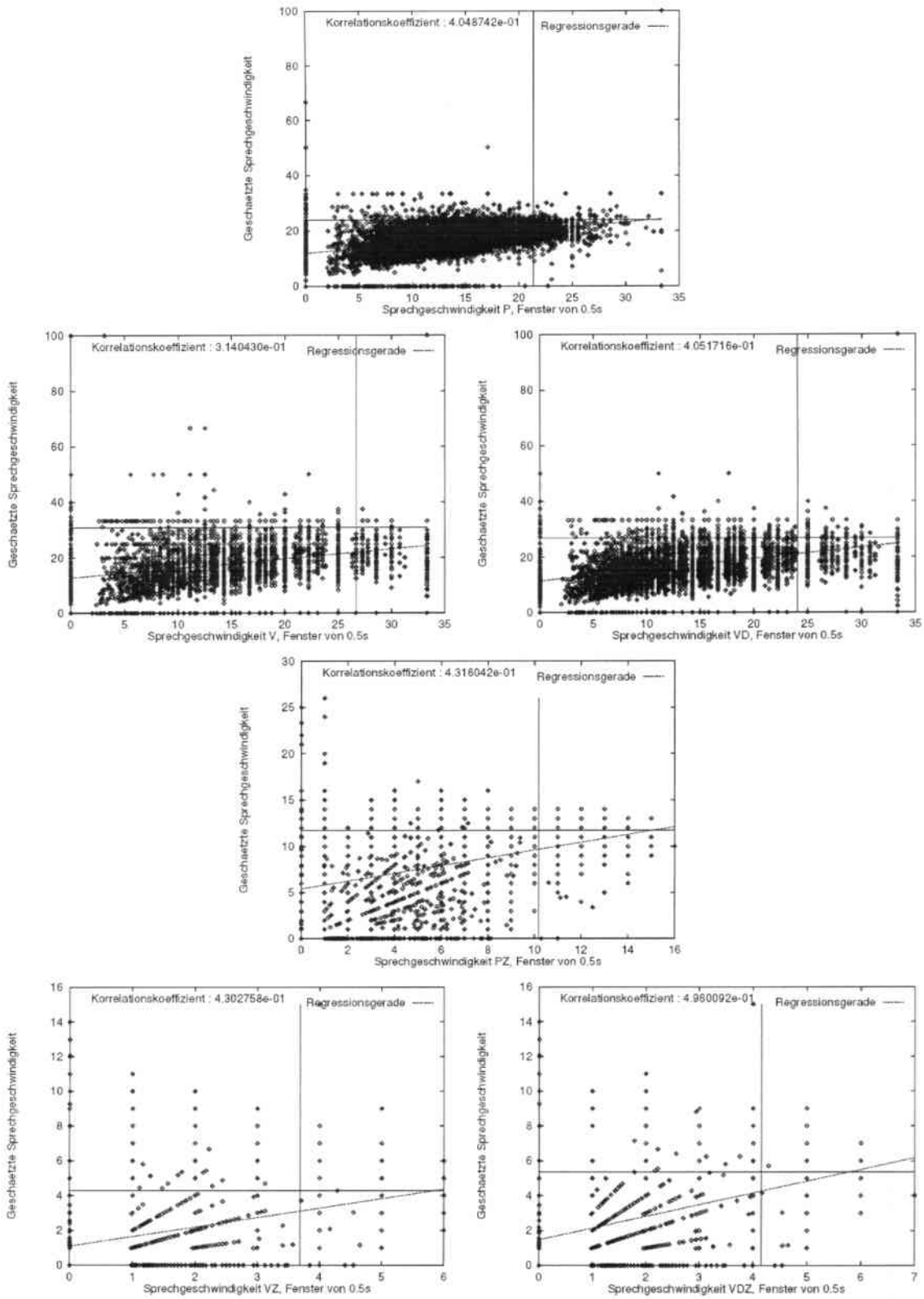


Abbildung 7: Scatterplots für die Schätzung von Sprechgeschwindigkeiten aufgrund der Hypothese des kontextunabhängigen Phonemerkenners auf Fenster von 0,5 Sekunden

5.3 Entropiebasierte Schätzung

5.3.1 Schätzung der Sprechgeschwindigkeit aufgrund des Entropieverlaufes

Gegeben sei eine endliche Menge von möglichen Versuchsausgängen $A_1, A_2, \ldots A_n$, die mit Wahrscheinlichkeiten $p_1, p_2, \ldots, p_n$ auftreten können ($\sum_{i=1}^{n} p_i = 1$ und $p_i \geq 0$ für alle i). Für die Unsicherheit eines einzelnen Ereignisses (bzw. Wahlfreiheit oder Information) führte Shannon ([15]) eine Maßeinheit ein, die S-Entropie, definiert als

$$H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

Für ein festes n hat die S-Entropie folgende Eigenschaften:

- $H \geq 0$ für alle $(p_1, p_2, \ldots, p_n)$.
- $H = 0$ genau dann, wenn alle p_i außer einem einzigen 0 sind, dieses eine p_i ist dann 1 (sichere Situation).
- Für ein gegebenes n erreicht H genau für die Gleichverteilung $(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ sein Maximum $(\log n)$ (unsichere Situation).
- H wächst bei jeder Veränderung der Wahrscheinlichkeiten $p_1, p_2, \ldots, p_n$ in Richtung Gleichverteilung.

Für $n = 2$ ist der Verlauf der S-Entropie H in Funktion der Wahrscheinlichkeiten $(p, p-1)$ in Abbildung 8 dargestellt.

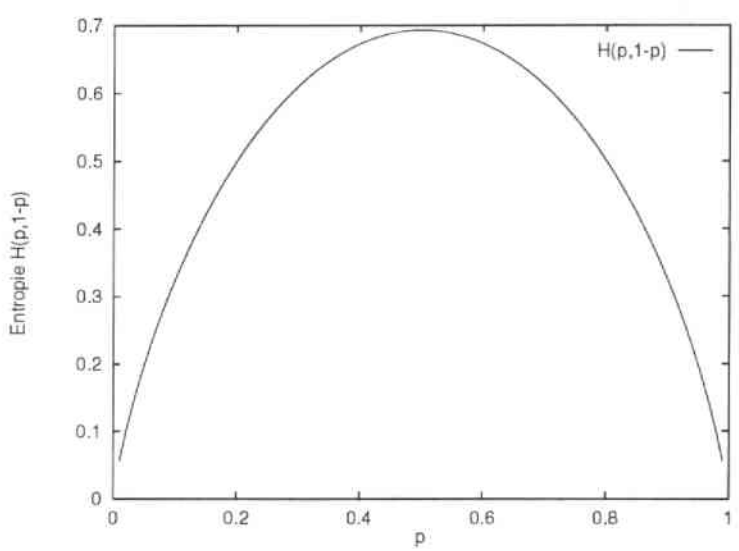


Abbildung 8: Entropieverlauf für 2 Ereignisse

Für den kontextunabhängigen Erkenner aus 3.3 modellieren die mittleren Zustände der Phonemmodelle (Abbildung 1, Teil m) die Akustik des mittleren Teils des Phonems ohne Koartikulationseffekte. Wie groß die akustische Ähnlichkeit eines gegebenen Frames zu den gelernten „reinen" Phonemen ist, spiegelt sich in den Werten der Emissionswahrscheinlichkeiten dieser mittleren Zustände wider. Wenn man alle diese Produktionwahrscheinlichkeiten durch ihre Gesamtsumme teilt, um die Normiertheitsbedingung $\sum_{i=1}^{n} p_i = 1$ zu erfüllen, kann man sie als einen Wahrscheinlichkeitsraum betrachten. Dann ist die S-Entropie pro Frame ein Maß für die Unsicherheit des Systems bei der Auswahl des reinen Phonems für das beobachtete akustische Ereignis. Genau berechnet sich die akustische Entropie als:

$$E(i) = -\sum_{j=1}^{N} \left(\frac{p_{ij}}{s_i}\right) \log \left(\frac{p_{ij}}{s_i}\right)$$

wobei:

p_{ij} = Wahrscheinlichkeit(Frame i | akustisches Modell j) (Emissionswahrscheinlichkeit des akustischen Modells j für ein Frame i)

$s_i = \sum_{j=1}^{N} p_{ij}$ (Summe der Emissionswahrscheinlichkeiten aller akustischen Modelle für das Frame i)

N = Anzahl der akustischen Modelle

ist.

Bei stärkerer Gleichverteilung (höherer Wert für die Entropie) kann man annehmen, daß keines der akustischen Modelle gut paßt, was eventuell einem Phonemübergang entspricht. Zu überprüfen ist also, ob der Entropieverlauf zur Schätzung der Sprechgeschwindigkeit geeignet ist.

Die Formel zur Schätzung der Sprechgeschwindigkeit aufgrund der Entropie lautet für ein Zeitfenster:

- SE = Anzahl lokaler Maxima der Entropie/Fensterlänge

Gemessen wurde die Korrelation der so gemessenen Sprechgeschwindigkeit SE mit der tatsächlichen berechnet als PZ (Anzahl Phonemanfänge/Fenster) und PDZ (Anzahl Phonemanfänge + Anzahl Diphtonganfänge/Fenster).

5.3.2 Wahl der akustischen Modelle

Als erstes wurde auf einem Teil der Trainingsmenge (426 Äußerungen von 50 Sprechern, insgesamt 1 Stunde und 2 Minuten) die Sprechgeschwindigkeit SE und PZ

und PDZ auf Fenster von 2 Sekunden gemessen. Es wurden 3 kontextunabhängige Phonemerkenner verwendet, die sich bei den akustischen Modellen unterscheiden: einer mit Anfang-, Mitte- und Ende-Modellen, einer mit Anfang- und Ende-Modellen und einer mit nur Mitte-Modellen (siehe 1).

Die Korrelationskoeffizienten ergaben:

Menge	Typ	Anfang, Mitte, Ende	Anfang, Ende	Mitte
gesamt	SE-PD	0,10	0,10	0,12
	SE-PDZ	0,11	0,09	0,12
schnell	SE-PD	0,11	0,14	0,26
	SE-PDZ	0,15	0,05	0,23

Wie erwartet hat man ein leicht besseres Ergebnis, wenn man nur die mittleren Modelle betrachtet.

5.3.3 Glättung

Als Beispiel betrachte man die Äußerung fdm1_mth1_tsponti3_fdm1_3_03 (aus der Test-, nicht aus der Trainingsmenge). Der Entropieverlauf und die entsprechenden Phonemgrenzen sind in Abbildung 9 gezeichnet.

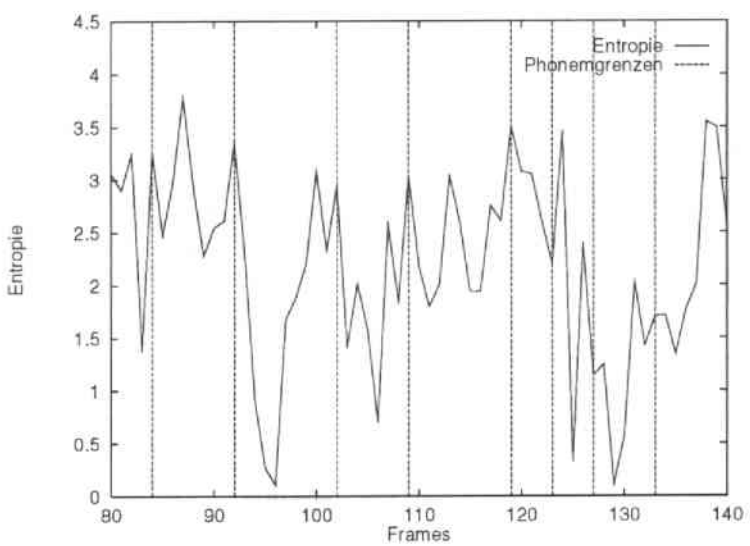


Abbildung 9: Entropieverlauf und Phonemgrenzen für die Frames 80 bis 140 der Äußerung fdm1_mth1_tsponti3_fdm1_3_03

Tatsächlich befinden sich lokale Maxima ungefähr bei den Phonemgrenzen. Man sieht allerdings, daß die lokalen Maxima viel öfter auftreten als die Phonemgenzen. Deshalb wurde der Entropieverlauf mit zwei Methoden geglättet:

$$\text{E\_3Glättung}(n) = \frac{0.5E(n-1) + E(n) + 0.5E(n+1)}{2}$$

$$\text{E\_5Glättung}(n) = \frac{0.3E(n-2) + 0.7E(n-1) + E(n) + 0.7E(n+1) + 0.3E(n+2)}{3}$$

Der geglättete Entropieverlauf ist in Abbildung 10 gezeichnet. Die Anzahl der Maxima ist jetzt der Anzahl der Phonemgrenzen ungefähr gleich.

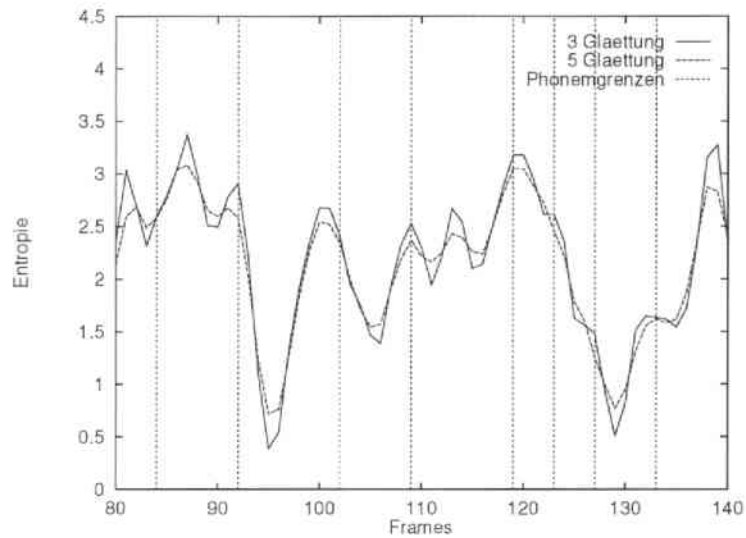


Abbildung 10: Geglätteter Entropieverlauf und Phonemgrenzen für die Äußerung fdm1_mth1_tsponti3_fdm1_3_03, Frames 80 bis 140

Auf den Testdaten wurden die Sprechgeschwindigkeiten gemessen und mit PZ und PDZ verglichen. Es wurde keine lineare Korrelation festgestellt (siehe Tabelle 12).

Korrelation	Fenster 0,5s	Fenster 1s	Fenster 2s
SE - PZ	0,00	-0,01	-0,02
SE - PDZ	-0,00	-0,02	-0,00
SE_3Glättung - PZ	0,04	0,04	0,03
SE_3Glättung - PDZ	0,03	0,03	0,02
SE_5Glättung - PZ	0,04	0,05	0,05
SE_5Glättung - PDZ	0,04	0,04	0,04

Tabelle 12: Korrelationskoeffizienten für die entropiebasierte Schätzung der Sprechgeschwindigkeit auf der gesamten Testmenge

5.3.4 Zusammenfassen in Phonemklassen

Durch Betrachten der genauen Score-Verteilungen für jedes einzelne Frame und die entsprechenden Entropiewerte bemerkt man, daß oft bestimmte Gruppen von Pho-

nemen bei einem Frame "ausschlagen". Es sind gerade die Gruppen der Phoneme, die untereinander öfter verwechselt werden. Zum Beispiel werden bei einem gesprochenen "A" sowohl das A-Modell, als auch die akustischen Modelle für AR, AHR und AH hohe Emissionsswahrscheinlichkeiten haben. Genauso für X auch K, R H +hLG oder für D auch B, P, T, G und K. Dies liegt daran, daß sich die Laute zum Teil sehr ähnlich sind oder von manchen Sprechern ähnlich gesprochen werden. Da die HMM-Modelle nicht diskriminativ gelernt werden, liegen sie näher "beieinander" und können nicht gut getrennt werden. In der Abbildung 11 ist für die Äußerung fdm1_mth1_tsponti3_fdm1_3_03, Frames 90 bis 104, ein Beispiel für die Score-Verteilung in jedem Frame gezeichnet.

Deshalb wurden Phoneme in Klassen zusammengefaßt, und ihre Emissionwahrscheinlichkeiten innerhalb der Klassen addiert. Man erhält einen Score pro Klasse. Für jedes Frame wurde wie oben die Entropie berechnet.

Es wurden folgende Klassenmengen gebildet:

Variante 1:

Klasse	Phoneme
MUELL	SIL +QK +hBR +hEH +hEM +hGH +hHM +hLG +hSM +nGN +nKL +nMK
FRIKATIVE	F V S Z SCH CH J X R H TS TSCH
PLOSIVE	P B T D K G
VOKALE	A AEH AH E E2 EH ER2 I IE O OE OEH ANG OH U UE UEH UH AR AEHR AHR AI AU EHR ER EU IR IHR OR OHR UR UEHR UHR
STIMMHAFTE KONSONANTEN	M N NG L R

Variante 2:

Klasse	Phoneme
MUELL	SIL +QK +hBR +hEH +hEM +hGH +hHM +hLG +hSM +nGN +nKL +nMK
FRIKATIVE	F V S Z SCH CH J X R H TS TSCH
PLOSIVE	P B T D K G
VOKALE und STIMMHAFTE KONSONANTEN	A AEH AH E E2 EH ER2 I IE O OE OEH ANG OH U UE UEH UH AR AEHR AHR AI AU EHR ER EU IR IHR OR OHR UR UEHR UHR M N NG L R

Es wurde verglichen, ob es günstiger ist, die MUELL-Modelle als eine Klasse zu betrachten oder sie zu ignorieren. Aus Abbildung 12 ist ersichtlich, daß der Entropie-

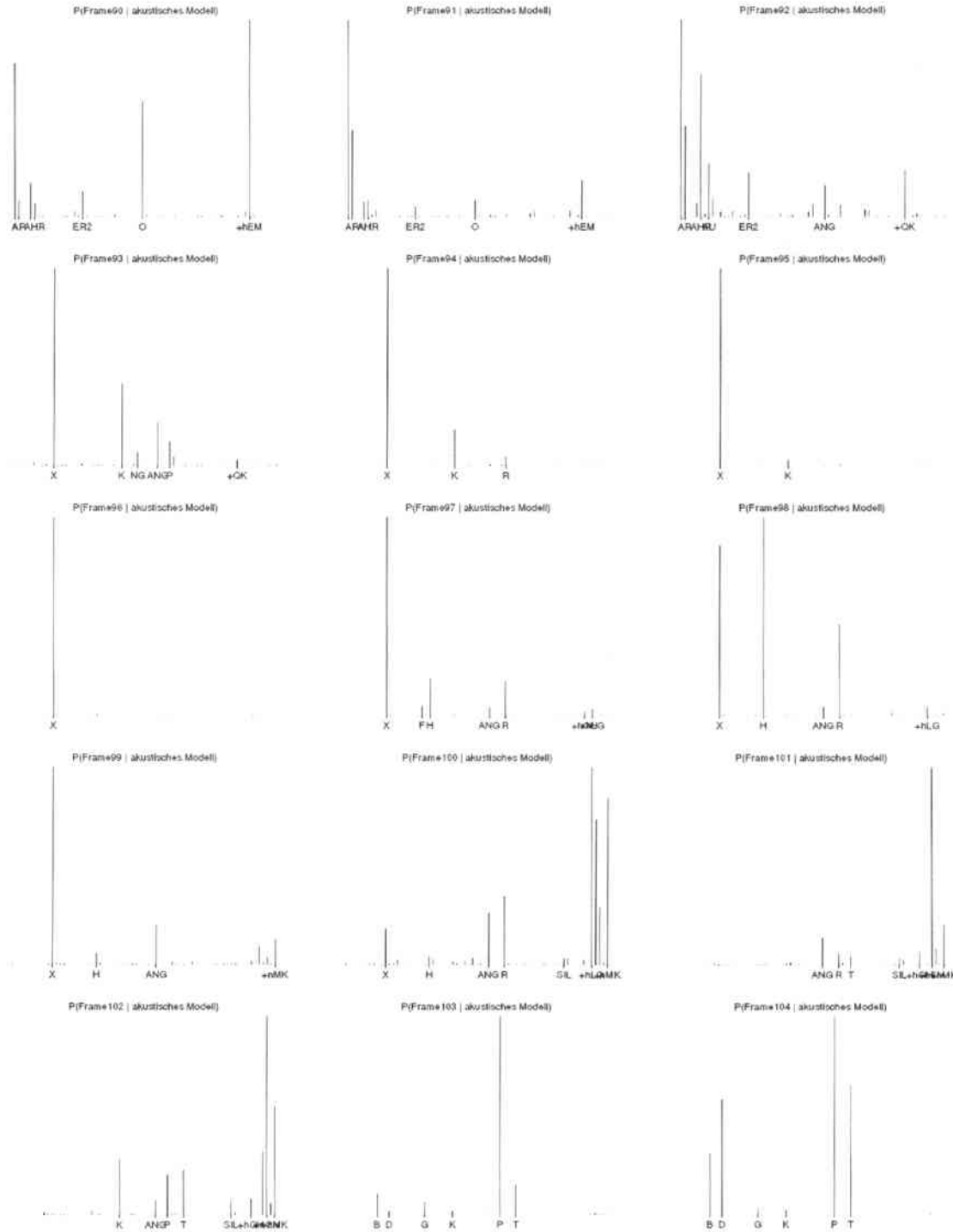


Abbildung 11: Verteilung der Emissionswahrscheinlichkeiten für die Frames 90 bis 104 der Äußerung fdm1_mth1_tsponti3_fdm1_3_03 („ach darf ich nich' so lange sprechen"). Von links nach rechts und oben nach unten sind folgende Phoneme den Frames 90 bis 104 zugeordnet: A X X X X X X X X X X D D D D.

verlauf in erwünschter Weise geglättet wird, wenn die MUELL-Modelle weggelassen werden. Da Geräusche und Stille vor den Schätzungen weggelassen wurden, ist es verständlich, daß sie die Schätzung verfälschen, wenn sie gut zu der Akustik passen.

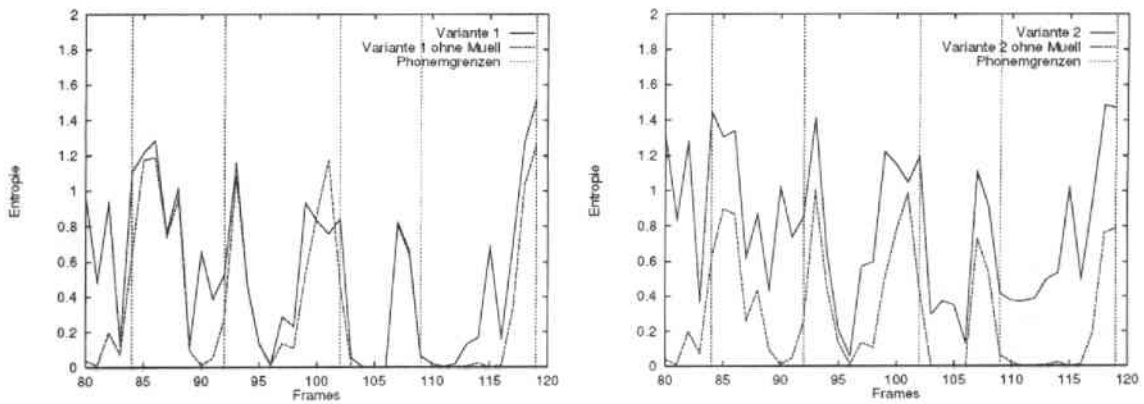


Abbildung 12: Vergleich zwischen dem Entropieverlauf für 5 Klassen (Variante1, links) bzw. 4 Klassen (Variante 2, rechts) mit und ohne MUELL (Geräusche und Stille)

Anschließend wurden auf den Testdaten die Korrelationskoeffizienten für die beiden Varianten ohne MUELL mit und ohne Glättung ermittelt (siehe 13).

Klassen	Korrelation	Fenster 0,5s	Fenster 1s	Fenster 2s
Variante 1 ohne Muell	SE - PZ	-0,11	-0,13	-0,14
	SE_3Glättung - PZ	0,14	0,15	0,15
	SE_5Glättung - PZ	0,01	0,03	0,02
Variante 2 ohne Muell	SE - PZ	-0,11	-0,15	-0,16
	SE_3Glättung - PZ	0,14	0,16	0,15
	SE_5Glättung	0,01	0,02	0,03

Tabelle 13: Korrelationskoeffizienten für die beiden Varianten der klassenbasierten Entropieschätzung ohne MUELL mit und ohne Glättung

Insgesamt sind die linearen Korrelationen höher als ohne Zusammenfassen in Klassen, aber immer noch unter 0,2. Eine 5-Glättung ist zu stark. Interessant ist, daß die Korrelationswerte ohne Glättung konsistent negative Werte ergeben, mit einer 3-Glättung positive. Die Ursache liegt in den akustischen Schwankungen innerhalb von längeren Phonemen. Diese führen zu kleinen Schwankungen der Entropie und dadurch zu vielen lokalen Maxima und zu hohen Werten für die Sprechgeschwindigkeit SE. Bei kurzen Phonemen besteht der Kern aus weniger Frames, kann also nicht so oft schwanken. Deshalb hat man bei einer tatsächlich niedrigen Sprechgeschwindigkeit, mit mehreren langen Phonemen, einen höheren Schätzwert SE als in Fenstern mit einer kleineren Sprechgeschwindigkeit. Eine 3-Glättung läßt gerade solche sehr feinen Schwankungen verschwinden.

Insgesamt ist klar, daß SE nicht eine Modalität darstellt, Sprechgeschwindigkeit gemessen als Anzahl der Phonemanfänge pro Sekunde zu approximieren. Der wich-

tigste Grund dafür ist der, daß die Schätzung von der Güte der akustischen Modelle abhängt. Mit perfekten Modellen bräuchte man aber keine Sprechgeschwindigkeitsschätzung mehr. Andererseits haben kleine akustische Variationen innerhalb von Phonemen einen großen (zumindest nicht klar vorhersehbaren) Einfluß auf den Entropieverlauf.

6 Zusammenfassung und Ausblick

In dieser Arbeit wurden mehrere Möglichkeiten untersucht, Sprechgeschwindigkeit für Spontansprache zu schätzen. Die Motivation dafür ist die Verbesserung der Erkennungsleistung von automatischen Spracherkennern mit Hilfe der Information über die Sprechgeschwindigkeit.

Zuerst wurden die Möglichkeiten besprochen, Sprechgeschwindigkeit zu messen. Wichtige Aspekte sind dabei die Wahl der Zeitintervalle, für die die Sprechgeschwindigkeit ermittelt wird, die Wahl der Maßeinheit und die Betrachtung von nichtsprachlichen Teilen (Stille und Geräuschen). Es wurde festgestellt, daß es im dem Fall, in dem überhaupt keine Vorinformation über die Spracheingabe vorhanden ist, am einfachsten und allgemeinsten ist, Sprechgeschwindigkeit für Zeitfenster konstanter Größe zu berechnen. Die Berechnung der Sprechgeschwindigkeit auf Phonemebene und Silbenebene wurde der Wort-Sprechgeschwindigkeit vorgezogen, da letztere die korrekten Wortgrenzen erfordert. Pausen und Geräusche wurden weggeschnitten, da sie nichts mit der *Sprech*geschwindigkeit zu tun haben.

Für die Approximation der Sprechgeschwindigkeit wurden unterschiedliche Berechnungsweisen betrachtet. Im Vordergrund stand dabei, daß sie einen Hinweis für Fehlerkennung bei schneller Sprechgeschwindigkeit geben. Es wurden folgende Berechnungsarten für die Sprechgeschwindigkeit aufgegriffen:

- Anzahl Phoneme / Gesamtdauer Phoneme
- Anzahl Vokale / Gesamtdauer Vokale
- Anzahl Phonemanfänge / Fensterlänge
- Anzahl Vokalanfänge / Fensterlänge

Auf einer Testmenge von insgesamt 66 Minuten, davon 44 Minuten reine Sprache, wurden zuerst die Phonemgrenzen automatisch mit einem *forced alignment* auf der korrekten Worttranskription durchgeführt. Diese Testmenge ist eine Teilmenge der GSST-Datenbasis und enthält spontane Dialoge zur Terminabsprache.

Die Äußerungen aus der Testmenge wurden dem Janus-Spracherkenner, der 1996 bei der Verbmobil-Evaluation eine Wortfehlerrate von 13,8% erzielte, zur Erkennung vorgelegt. Daraufhin wurden die Worte aus der Testmenge in zwei Klassen aufgeteilt: eine Klasse der korrekt erkannten und eine Klasse der falsch erkannten Worte. Es wurden für jedes Wort die Sprechgeschwindigkeiten mit den unterschiedlichen Formeln berechnet und deren Verteilung betrachtet. Die Mittelwerte waren insgesamt höher auf den falsch erkannten Worten als auf den korrekt erkannten, so daß ein Zusammenhang zwischen hohen Werten der Sprechgeschwindigkeit und Fehlerkennung festgestellt wurde.

Um die Sprechgeschwindigkeit zu schätzen, wurde zuerst der Verbmobil-Erkenner eingesetzt, und eine Segmentierung aufgrund seiner Hypothese mit einem *forced alignment* durchgeführt, anhand der dann die Berechnung der Sprechgeschwindigkeit erfolgte. In den Arbeiten von Mirghafori und Siegler wurden damit sehr gute Schätzungen auf Daten aus der WSJ-Datenbasis erhalten. Auf der Testmenge (Spontansprache) wurden in der vorliegenen Arbeit die Ergebnisse im Prinzip bestätigt. Es wurden für Fenster von 1 Sekunde hohe lineare Korrelationen von 0,83 auf den gesamten Daten gemessen. Diese Schätzungsweise ist allerdings zeitaufwendig und gerade in kritischen Fällen (schnell gesprochene Sätze) unzuverlässig, da der Korrelationskoeffizient auf den schnellsten 5% der Zeitfenster nur noch 0,31 beträgt.

Danach wurde eine Methode ausprobiert, die Anzahl der Phonemübergänge aus dem Entropieverlauf der akustischen Emissionswahrscheinlichkeiten eines kontextunabhängigen Spracherkenners zu ermitteln. Es wurden die Maxima der akustischen Entropie gezählt, und mit der Anzahl der Phonemanfänge korreliert. Dabei wurden unterschiedliche Glättungsvarianten für die Entropie ausprobiert. Es konnte keine lineare Korrelation festgestellt werden. Danach wurden diejenigen Phoneme, die oft untereinander verwechselt werden, in Klassen zusammengefaßt. Es wurden wiederum mehrere Varianten ausprobiert, wobei der beste Korrelationswert bei 0,16 lag.

Deshalb wurde die Hypothese eines kleinen, schnellen, kontextunabhängigen Phonemerkenners zur Schätzung betrachtet. Es wurde versucht, damit sowohl die Phonem- als auch die Vokalgeschwindigkeit zu approximieren. Die Korrelation mit der tatsächlichen Sprechgeschwindigkeit betrug zwischen 0,3 und 0,5 auf der gesamten Testmenge, und bis zu 0,3 auf den schnellsten 5% der Werte. Die Schätzung ist besser als aufgrund der Entropie, weil die akustischen Schwankungen innerhalb der Phoneme bei der Suche durch eine Gesamtoptimierung über mehrere Frames kompensiert werden. Allerdings ist der kontextunabhängige Phonemerkenner zu ungenau, um eine wirklich gute Schätzung der Sprechgeschwindigkeit aufgrund seiner Hypothese zu ermöglichen.

Schlußfolgernd kann man sagen, daß alle in dieser Arbeit betrachteten Methoden abhängig von der Güte der verwendeten akustischen Modelle oder von einem Spracherkenner sind, und daß sie deshalb keine befriedigenden Schätzungen der Sprechgeschwindigkeit ergeben.

Um bei der Schätzung der Sprechgeschwindigkeit vom Spracherkenner unabhängig zu sein, ist es sinnvoll, sie direkt auf dem akustischen Signal vorzunehmen. Eine Möglichkeit dafür ist die Detektierung von Vokalen aufgrund der Energie und *Zero Crossing Rate*, Ansätze die in Berkeley und München gerade ausprobiert werden. Andere Möglichkeiten sind die Analyse des Verlaufs und der Variationen der Spektralkoeffizienten. Denkbar ist auch der Einsatz von neuronalen Netzen zur Detektierung von Phonemanfängen wie in [10, 18] angedeutet, oder direkt zur Approximation der Sprechgeschwindigkeit.

Literatur

[1] A. Anastakos, R. Schwartz. H. Shu: *Duration Modeling in Large Vocabulary Speech Recognition*, Proc. ICASSP 1995

[2] M.Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal: *The Karlsruhe-Verbmobil Speech Recognition Engine*, Proc. ICASSP 1997

[3] W.S. Humphrey: *A Discipline for Software Engineering*, Addison-Wesley 1995

[4] M. Finke, I. Rogina: *Wide Context Acoustic Modeling in Read vs. Spontaneous Speech*, Proc. ICASSP 1997

[5] W. Gellert: *Kleine Enzyklopädie Mathematik*, Bibliographisches Institut, Leipzig, 1986

[6] K. Hinderer: *Stochastik für Informatiker und Ingenieure*, Skriptum, Institut für Mathematische Statistik, Universität Karlsruhe, 1989

[7] T. Kemp: *Regelbasiert generierte Aussprachevarianten für Spontansprache*, Proc. KONVENS 96

[8] L. Lee, R. C. Rose: *Speaker Normalization using Efficient Frequency Warping Procedures*, Proc. ICASSP 1996

[9] N. Mirghafori, E. Fosler, N. Morgan: *Making Automatic Speech Recognition More Robust to Fast Fast Speech*, December 1995, Technical Report ICSI Berkeley

[10] K. Paschen: *Lookahead mit Neuronalen Netzen in JANUS*, Studienarbeit ILKD, 1997

[11] D. Pötschke, F. Sobik: *Mathematische Informationstheorie*, Akademie-Verlag, Berlin 1980

[12] L.R. Rabiner, R.W. Schafer: *Digital Processing of Speech Signals*, Prentice-Hall 1978 A

[13] T. Schaaf, T. Kemp: *Confidence Measures for Spontaneous Speech Recognition*, Proc. ICASSP 1997

[14] E. G. Schukat-Talamazzini: *Automatische Spracherkennung*, Vieweg, Braunschweig/Wiesbaden 1995

[15] C. E. Shannon, W. Weaver: *Mathematische Grundlagen der Informationstheorie*, Oldenbourg Verlag, München 1976

[16] M.-A. Siegler, R. M. Stern: *On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems*, Proc. ICASSP 1995

[17] M.-A. Siegler: *Measuring and Compensating for the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition*, MS Thesis, School of Computer Science, Carnegie Mellon University, 1995

[18] J. Verhasselt, J.-P. Martens: *A Fast and Reliable Rate of Speech Detector*, Proc. KSLP 1996

[19] M. Westphal: persönliche Kommunikation

[20] M. Westphal, P. Zhan: *Speaker Normalization Based on Frequency Warping*, Proc. ICASSP 1997

A Histogramme verschiedener tatsächlicher Sprechgeschwindigkeiten

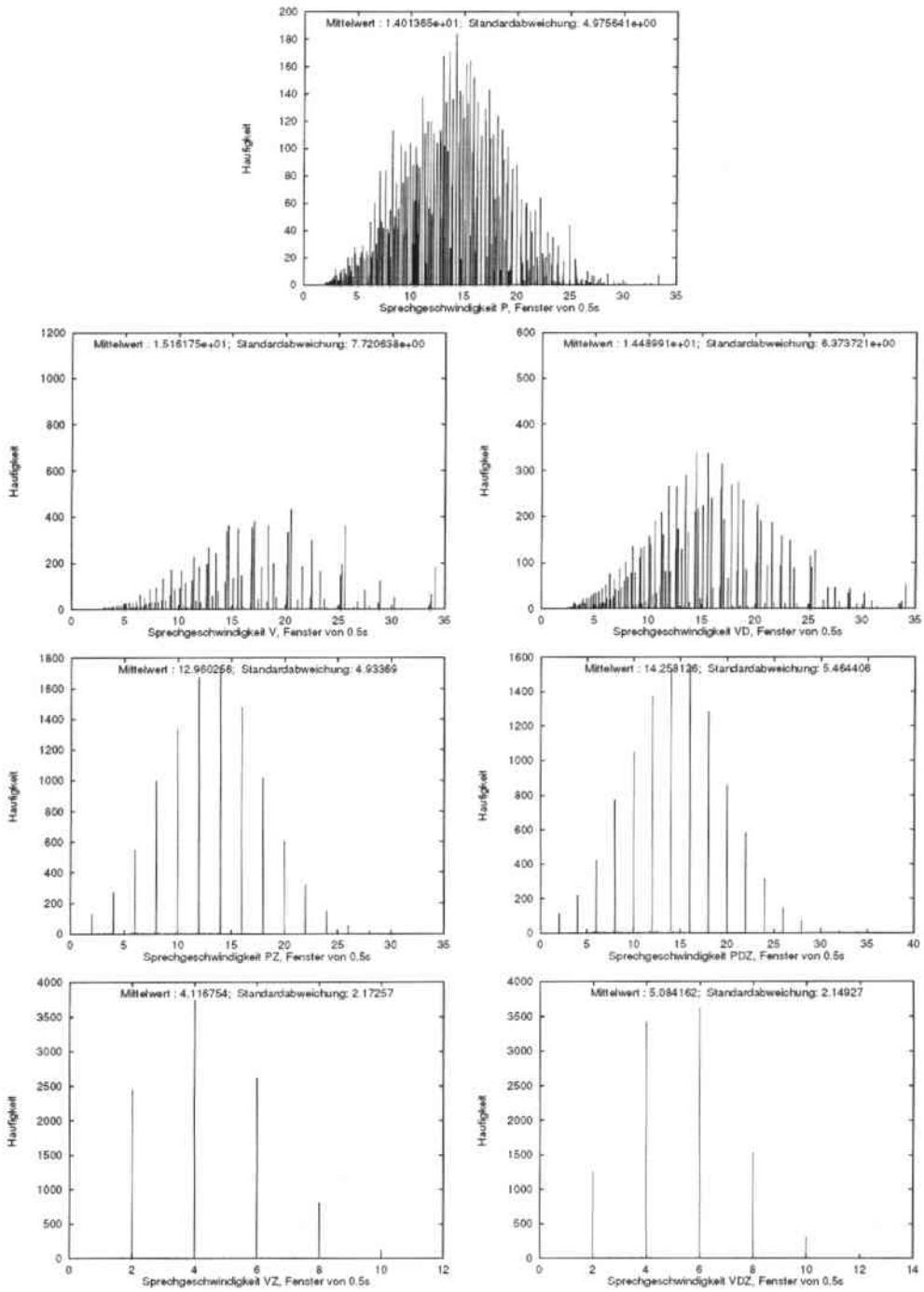


Abbildung 13: Histogramm für die tatsächlichen Sprechgeschwindigkeiten P, V, VD, PZ, PDZ, VZ und VDZ für Fenster der Länge 0,5 Sekunden auf der gesamten Testmenge

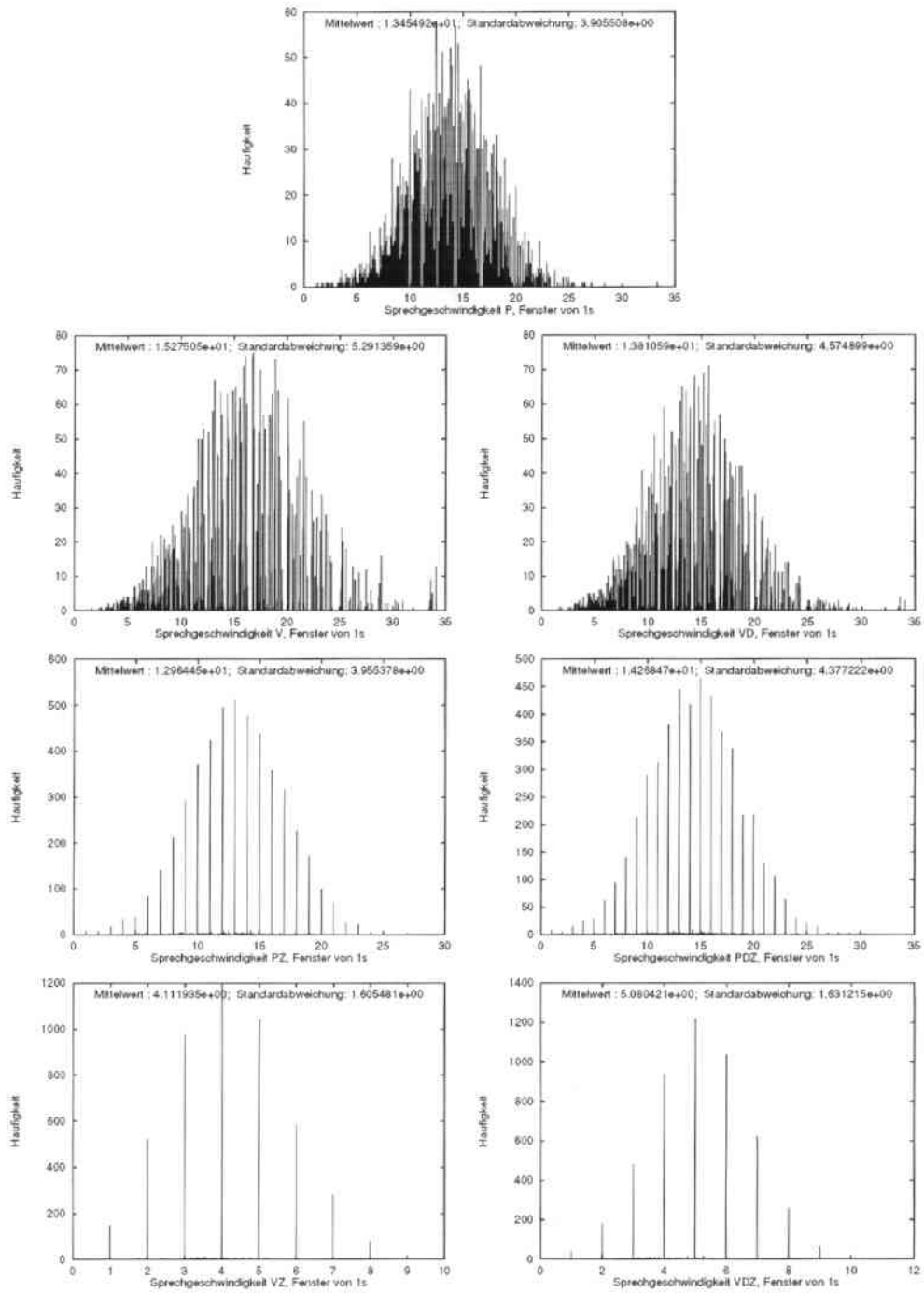


Abbildung 14: Histogramm für die tatsächlichen Sprechgeschwindigkeiten P, V, VD, PZ, PDZ, VZ und VDZ für Fenster der Länge 1 Sekunde auf der gesamten Testmenge

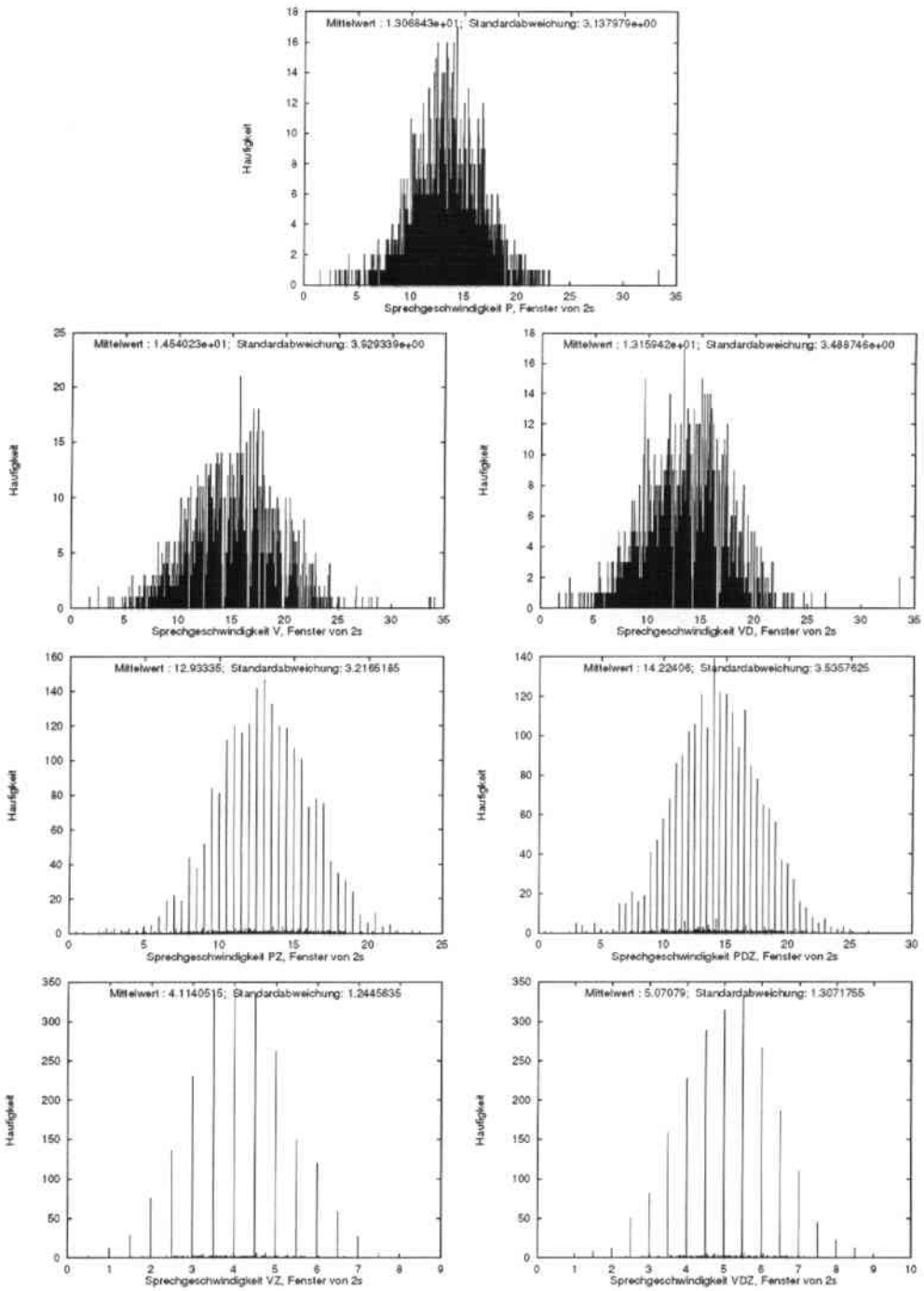


Abbildung 15: Histogramm für die tatsächlichen Sprechgeschwindigkeiten P, V, VD, PZ, PDZ, VZ und VDZ für Fenster der Länge 2 Sekunden auf der gesamten Testmenge