BMBF

Verbmobil
Verbundvorhaben

Das Münchener AUtomatische Segmentationssystem (MAUS)

Andreas Kipp

M.-B. Wesenick

Ludwig-Maximilians-Universität München

Vm Memo Nr. 95
November 1995

November 1995

Andreas Kipp
Maria-Barbara Wesenick

Institut für Phonetik und Sprachliche Kommunikation
Ludwig-Maximilians-Universität München
Schellingstraße 3/II
80799 München

Tel.: (089) 2180 - 2810

e-mail: kip@sun1.phonetik.uni-muenchen.de

Gehört zum Antragsabschnitt: 14 VERBMOBIL/PHONDAT

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 102 L/4 gefördert. Die Verantwortung für den Inhalt liegt bei den Autoren.

Inhaltsverzeichnis

1 Einführung 4

2 Datenbank und Lexikon 5

3 Die einzelnen Verarbeitungskomponenten von MAUS 6

3.1 Vorverarbeitung 6

3.2 Statistische Modelle 6

3.3 Variantengenerator 7

3.3.1 Korpus mit Ausspracheregeln 7

3.3.2 Erzeugen eines Variantengraphen 7

3.4 Viterbi-Suche 9

3.5 Refinement 9

4 Evaluation 10

4.1 Allgemeine Problematik 10

4.2 Ergebnisse 10

4.2.1 Sprachmaterial 10

4.2.2 Übereinstimmung der Label 10

4.2.3 Übereinstimmung der Segmentgrenzen 11

5 Zusammenfassung 13

6 Literatur 13

Das Münchener Automatische Segmentationssystem (MAUS)

Im Rahmen von VM-TP 14.7 wird im Institut für Phonetik und Sprachliche Kommunikation München ein automatisches System zur Segmentation von gesprochener Sprache entwickelt. Es handelt sich um ein hybrides statistisches und regelbasiertes System, das als Eingabe das Zeitsignal und die orthographische Repräsentation einer Äußerung benötigt. Ergebnis ist eine phonetische Transkription der Äußerung und die entsprechende Segmentation des Sprachsignals. Die Architektur des Systems und seine Arbeitsweise, sowie eine erste Evaluierung seiner Analyseleistung werden im folgenden beschrieben.

1 Einführung

Ziel des Teilprojekts ist es, ein System zu entwickeln, mit dem große Mengen von Sprachmaterial automatisch segmentiert werden können. Voraussetzung ist, daß für die zu segmentierende Äußerung neben den Signaldaten eine entsprechende orthographische Form vorliegt. Das Münchener Automatische Segmentationssystem (MAUS) liefert nach der Bearbeitung eine breite phonematische Transkription der Äußerung und die jeweiligen Segmentgrenzen.[1]

Anwendung finden soll das System in Bereichen, in denen umfangreiches segmentiertes Sprachmaterial benötigt wird. Dazu gehört u. a. das Gebiet der automatischen Spracherkennung, wenn es darum geht, statistische Modelle zu initialisieren und trainieren. Auf der Grundlage von sehr großen Sprachdatenbanken mit segmentiertem Material können aber auch großangelegte empirische phonetische/phonologische Untersuchungen durchgeführt und beispielsweise Wortformenstatistiken berechnet werden.

Segmentationen werden gewöhnlich manuell erstellt, was aber in Anbetracht sehr umfangreicher Sprachdatenbanken wegen des äußerst großen Zeitaufwands nahezu unmöglich ist. Ein System, das ausreichend zuverlässig automatische Segmentationen herstellen kann, ist demnach wünschenswert.

MAUS (Schema s. Abbildung 1) basiert auf Hidden Markov Modellen (HMM) für einzelne Phoneme des Deutschen (s. Abschnitt 3.2) und verarbeitet im Gegensatz zu gängigen Spracherkennungssystemen in einer regelbasierten Komponente zusätzlich Wissen über phonetische Prozesse im Deutschen. Über die Orthographie kann die kanonische Form einer Äußerung in einem Lexikon aufgefunden werden.[2] Diese phonetische Form wird der regelbasierten phonetischen Komponente zugeführt, die zur kanonischen Form der Äußerung eine Vielzahl von hypothetischen Aussprachevarianten in Form eines gerichteten Graphen erzeugt (s. Abschnitt 3.3). Die Knoten des Graphen repräsentieren statistische Modelle, Kanten stehen für die erlaubten Übergänge. Eine ausführliche Beschreibung des Regelsystems zur Generierung der Aussprachevarianten findet sich in [5].

Aus dem akustischen Sprachsignal werden in einer Vorverarbeitungskomponente relevante akustische Merkmale extrahiert (s. Abschnitt 3.1). Mittels Viterbi-Alignment wird anschließend derjenige Pfad durch den Variantengraphen ausgewählt, der im statistischen Sinn optimal auf das akustische Signal paßt (s. Abschnitt 3.4). Die Symbole der gewählten Modelle werden den entsprechenden Abschnitten des Signals zugeordnet, so daß am Ende dieses Analyseschritts für jedes gefundene Segment des Signals die Daten *Segmentbeginn*, *Segmentende* und *phonetisches Symbol*

1. Zur Terminologie: Eine Transkription mit zugehörigen Segmentgrenzen wird zusammengefaßt als *Segmentation* bezeichnet.
2. Zur Terminologie: Unter *kanonischer Form* ist diejenige Ausspracheform einer Äußerung zu verstehen, die als einzige Standardform aus mehreren möglichen Zitierformen (Standardaussprache isolierter Wörter) arbiträr ausgewählt wurde. Sie wird mit Hilfe von Symbolen eines phonetischen Alphabets notiert.

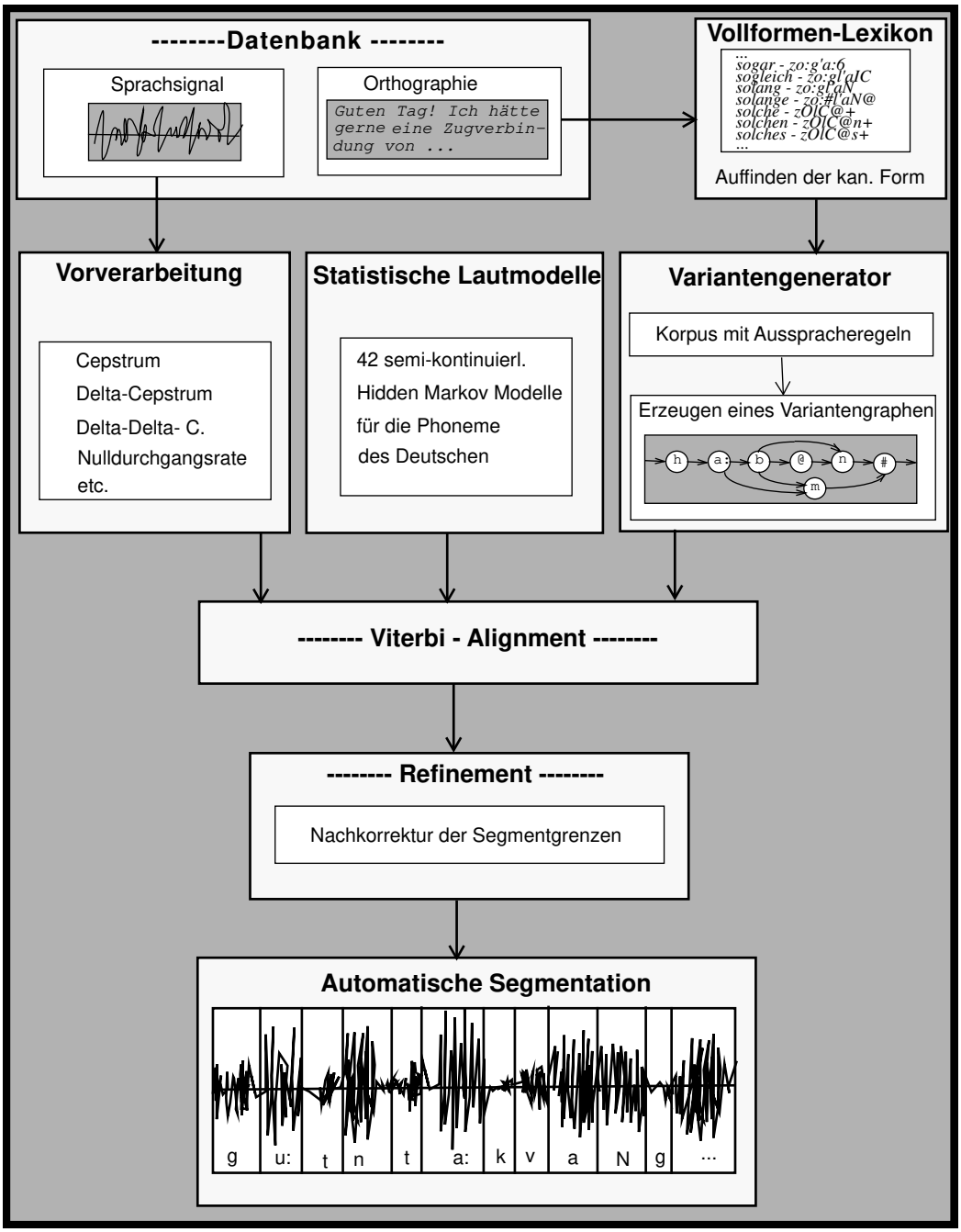

Abbildung 1: Schema des Gesamtsystems

vorliegen. In einer letzten Komponente des Systems (Refinement) werden die so errechneten Segmentgrenzen mit signalnahen Verfahren korrigiert und verfeinert (s. Abschnitt 3.5).

2 Datenbank und Lexikon

Bei Fertigstellung des Systems werden als Eingabedaten lediglich die Signaldaten der zu analysierenden Äußerung und Informationen über das Gesprochene in Form einer orthographischen

Repräsentation benötigt. Über ein Vollformen-Lexikon wird auf die kanonischen Formen der in der Äußerung enthaltenen Wörter zugegriffen, welche zu einer zusammengesetzten kanonischen Form der Gesamtäußerung aneinandergefügt und anschließend an den Variantengenerator weitergeleitet werden. Die akustischen Daten werden der Vorverarbeitungskomponente zur Feature-Extraktion zugeführt.

Da zum gegenwärtigen Zeitpunkt noch kein genügend umfangreiches Vollformen-Lexikon zur Verfügung steht, arbeiten wir vorerst mit Daten, die im Phondat-Format vorliegen, welches neben akustischen Daten und Orthographie auch die kanonischen Formen der Äußerung im Datei-Header enthält. Die kanonischen Formen werden dem Variantengenerator zugeführt.

Bisherige Analysen sind zunächst auf Daten des Phondat-II Zugauskunft-Korpus beschränkt worden, für das nun eine vollständige automatische Segmentation vorliegt. Da dieses Sprachdatenkorpus mehrfach manuell segmentiert wurde, ist eine aussagekräftige Evaluation der Analyseleistung von MAUS durchführbar, die für die Weiterentwicklung Voraussetzung ist.

Momentan werden Softwareanpassungen für die automatische Analyse von VERBMOBIL-Sprachdaten vorgenommen.

3 Die einzelnen Verarbeitungskomponenten von MAUS

3.1 Vorverarbeitung

Aus dem digitalen Sprachsignal (16 kHz Abtastrate) werden in einem 20 ms Fensterbereich (Hamming Window) alle 10 ms Feature-Vektoren bestehend aus folgenden Parametern extrahiert:

- 10 Cepstral-Koeffizienten mit 1. und 2. Ableitung
- normalisierte Energie und 1. Ableitung
- Nulldurchgangsrate und 1. Ableitung

3.2 Statistische Modelle

Für folgende Lauteinheiten sind statistische Modelle errechnet worden:

- Vollvokale: `a, a:, e:, I, i:, O, o:, U, u:, E, E:, 9, 2:, Y, y:`
- reduzierte Vokale: `@, 6`
- Diphthonge: `aI, aU, OY`
- Plosive: `p, b, t, d, k, g, Q`
- Frikative: `f, v, s, z, S, Z, C, j, x, h`
- Nasale: `m, n, N`
- Lateral, Vibrant: `l, r`

Dazu kommen zwei Pausenmodelle.

Bei den statistischen Modellen handelt es sich um kontextfreie semikontinuierliche Hidden Markov Modelle (scHMM) (nach [2]). Die oben genannten Merkmale werden in 5 getrennten Kodebüchern mit variierender Prototypenzahl (16 - 256) modelliert. Die Emmissionswahrscheinlichkeit berechnet sich als Produktkode der einzelnen Kodebücher. Jedes HMM kann 3 - 6 Zustände enthalten. Die Modelle wurden mit handsegmentiertem Material von 12 Sprechern ('SPICOS'-Material, ca. 2400 Äußerungen) initialisiert und mit segmental-k-means trainiert. Ein weiteres unüberwachtes Nachtraining mit dem PhonDat-I Korpus (201 Sprecher, 21681 Äußerungen) wurde zwar durchgeführt, da es aber aber keine weitere Verbesserung der Segmentierungsleistung brachte, werden die Modelle ohne Nachtraining verwendet.

3.3 Variantengenerator

3.3.1 Korpus mit Ausspracheregeln

Das Korpus besteht in seiner gegenwärtigen Form aus ca. 1550 Regeln. Das Symbolinventar umfaßt die o.g. phonetischen Symbole, für die ein HMM vorliegt (außer Pausen) und die folgenden Sonderzeichen:

`!v`	Vokal
`!K`	Konsonant
`!N`	Nasal
`#`	Wortgrenze
`&`	arbiträre Wortgrenze

Eine Regel r_i, i = 0...N besteht aus einem rechten und einem linken Teil, die durch ">" getrennt sind. Der linke Teil enthält einen Symbolstring $\mathbf{a_i} = <a_i(0), ... a_i(K_i - 1)>$, der einem Teilstring der kanonischen Form entsprechen muß. Auf der rechten Seite steht ein Symbolstring $\mathbf{b_i} = <a_i(0), ... a_i(K_i - 1)>$, der den variierten Symbolstring a_i der kanonischen Form repräsentiert. $a_i(k)$ und $b_i(l)$, $k = 0...K_i$, $l = 0...L_i$ sind Symbole des phonetischen SAM-Alphabets und die oben aufgeführten Sonderzeichen, ausgenommen "&" für die rechte Seite. Aus Verarbeitungsgründen wird jeder Regel eine einstellige Ziffer (default: 1) vorangestellt. Einige Beispiele zur Verdeutlichung:

`1nf>mf`	Die Symbole `nf` können durch `mf` ersetzt werden.
`1#pf>#f`	`pf` am Wortanfang kann durch `f` ersetzt werden.
`1p#j>p#C`	An Wortgrenzen kann `pj` zu `pC` werden.
`1g@n#>gN#`	Am Wortende kann `g@n` durch gN ersetzt werden.
`1t#t>&t`	Treffen zwei `t` an einer Wortgrenze aufeinander, können sie durch ein einziges ersetzt werden. Die Wortgrenze ist dann arbiträr und wird konventionsgemäß vor das betreffende Segment gesetzt.
`1!vtp>!vQp`	`tp` kann nach Vokal zu `Qp` werden.
`1!vr!K#>!v6!K#`	`r` kann am Wortende nach Vokal und vor Konsonant vokalisiert und durch `6` ersetzt werden.

Mit den Regeln können von der kanonischen Form einer Äußerung eine Vielzahl von möglichen Aussprachevarianten auch über Wortgrenzen hinweg abgeleitet werden. In einer früheren Version von MAUS (siehe auch [4]) wurden die Varianten als Listen gespeichert. Ein entscheidender Fortschritt ist die jetzige Repräsentation der kanonischen Form und ihrer Varianten in Form eines gerichteten Graphen. Dadurch wurde die Handhabung der großen Menge von Varianten enorm erleichtert, die Viterbi-Suche effizienter und wesentlich schneller.

3.3.2 Erzeugen eines Variantengraphen

Zur Erzeugung des Graphen wird zunächst die kanonische Form einer Äußerung als Graph dargestellt. In diesem einfachen Graph haben alle Knoten genau einen Nachfolger (bis auf den letzten). Der Graph besitzt genau einen Pfad vom Anfangs- zum Endknoten, entlang dessen die Lautsymbolfolge der kanonischen Form emittiert wird. Die kanonische Form des Worts *haben* ist als Beispiel in Abbildung 2 als Graph dargestellt.

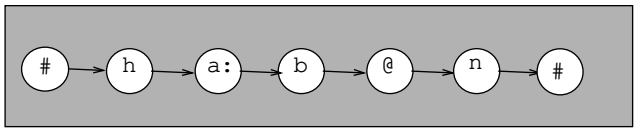

Abbildung 2: Darstellung einer kanonischen Form als gerichteter Graph

Nun wird für jede Regel, die auf die kanonische Form anwendbar ist, ein alternativer Pfad im Graph erzeugt. Auf das Beispielwort *haben* passen folgende Regeln:

```
1b@n#>bn#
1b@n#>bm#
1b@n#>m#
```

Der komplexe Graph der kanonischen Form mit Varianten für das Beispiel ist in der folgenden Abbildung 3 zu sehen.

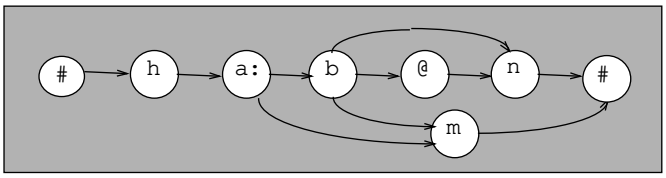

Abbildung 3: die kanonische Form mit Alternativpfaden

Um die Forderung nach einer minimalen Anzahl von Knoten zu erfüllen, werden für Symbole, die am Anfang und am Ende der rechten und der linken Seite identisch sind, keine neuen Knoten eingefügt. Abbildung 4 zeigt den Algorithmus zur Generierung des Variantengraphen in in einer Pseudo-Programmiersprache.

```
for i=0...N-1
   if the graph G(0) contains a node sequence n_a which emits
   a_I then:
      if L_i-n_i-m_i > 0 then
         add a node sequence n_b of length L_i-n_i-m_i emitting
         the symbols b_i(l),l=n_i...L_i-m_i-1;
         mark first node of n_b as start node N_start and last
         node of n_b as end node N_end of alternative path
      else mark the n_i th node of n_a as N_start and the L_i-m_ith
         as N_end (if either n_i=0 or m_i=0 N_start or N_end are
         undefined and not required in later pocessing)
      endif
      if n_i>0 then
         skip n_i-1 nodes from the beginning of n_a and add a
         transition from that node to N_start
      else keep in memory that transitions from all prede-
         cessors of the first node of n_a to N_start have to be
         inserted (pending transitions)
      endif
      if m_i>0 then
         skip back m_i-1 nodes from the end of n_a and add a
         transition from N_end to that node
      else keep in memory that transitions from N_end to all
         successors of the last node of n_a have to be
         inserted (pending transitions)
      endif
   endif
end for
repeat
   add pending transitions from inserted nodes to successors
   of nodes in G(0) (This may increase the number of prede-
   cessors of other nodes in G(0) and introduce new pending
   transitions);
   add pending trasitions from predecessor nodes in G(0) to
   inserted nodes (This may increase the number of predeces-
   sors of other nodes in G(0) and introduce new pending
   transitions);
until no more transitions have to be inserted
```

Abbildung 4: Algorithmus zur Generierung des Variantengraphen

3.4 Viterbi-Suche

Die Zuordnung zwischen Variantengraph und Zeitsignal erfolgt in einer HMM-basierten Stufe durch eine durch den Graphen eingeschränkte, datengesteuerte Viterbi-Suche. Die im Graphen enthaltene, im statistischen Sinn wahrscheinlichste Variante wird ermittelt und ein Alignment zwischen dieser Variante und dem Zeitsignal hergestellt. Dazu werden die oben beschriebenen scHMM verwendet.

Jedes Lautsymbol des Variantengraphen entspricht einem scHMM. Zunächst werden die scHMM der Anfangsknoten und ein Pausen-scHMM gestartet. Bei Erreichen des Endzustands eines HMMs werden ausschließlich diejenigen Nachfolge-HMM neu gestartet, die durch die im Variantengraphen festgelegten Transitionen des aktiven scHMMs bestimmt werden. Auf diese Weise wird die Suche ohne redundante Schritte auf die im Graph enthaltenen Varianten beschränkt. Die Etikettierung der Segmente und die dazugehörigen zeitlichen Markierungen werden durch Zurückverfolgen des Suchpfades (Backtracking) ermittelt.

Ergebnis dieses Analyseschritts ist eine vorläufige Segmentation der Äußerung, die als Liste aller ermittelten Segmente mit ihrem Label, dem Anfangssample und der Dauer dargestellt wird. Die folgende Tabelle 1 zeigt ein Segmentationsbeispiel:

Anfang	Dauer	Label
7203	6500	g
13704	1237	u:
14942	784	t
15727	1329	@
17057	548	n
17606	1391	t
18998	3329	a:
22328	3043	k
25372	2212	v
27585	1421	a
29007	1973	n
30981	300	g
31282	1260	e:
32543	817	p
33361	1302	m
34664	1879	O6
36544	1120	N
37665	1790	f
39456	1385	o:6
...		

Tabelle 1: Beispiel für die automatische Segmentation einer Äußerung

3.5 Refinement

In dieser letzten Stufe wird die vorläufige Segmentation, die aus dem Viterbi-Search hervorgeht, mit signalnahen Methoden verfeinert. So werden beispielsweise die errechneten Grenzen auf den nächstliegenden Nulldurchgang des Signals verschoben oder Korrekturen der Anfangsgrenzen von initialen Plosiven vorgenommen.

Es ist geplant, diese Stufe weiter auszubauen, um nach der Viterbi-Stufe Signalabschnitte gezielt zu behandeln, von denen man aus der Evaluation weiß, daß sie durch signalnahe Methoden verbessert werden können.

4 Evaluation

4.1 Allgemeine Problematik

Um quantitative Aussagen über die Analyseleistung von MAUS machen zu können, ist ein automatisches Verfahren für die Evaluation ausgearbeitet worden. Dabei stellte sich zunächst die Frage nach einer geeigneten Referenzsegmentation, auf deren Grundlage die automatische Segmentation beurteilt werden kann.

Die manuelle Analyse von Sprachsignalen unterliegt der subjektiven Beurteilung durch den menschlichen Segmentierer. Selbst gute Segmentationen desselben Sprachsignals, die von unterschiedlichen Phonetikern erstellt wurden, unterscheiden sich voneinander, ohne daß eine "richtiger" wäre als die andere (siehe auch [1]). Man kann demnach bei einer Abweichung der automatischen Segmentation von einer bestimmten manuell erstellten nicht grundsätzlich behaupten, die automatische Segmentation sei im betreffenden Fall nicht korrekt.

Unsere Evaluation bezieht sich entsprechend nicht auf eine bestimmete Referenzsegmentation, sondern auf die Segmentationen mehrerer menschlicher Segmentierer. Dazu verwenden wir Sprachmaterial, das von mehreren Segmentierern bearbeitet worden ist, und untersuchen die Abweichungen in den Segmentationen der menschlichen Segmentierer untereinander in Bezug auf Abweichungen von gewählten Segmentgrenzen und Transkriptionssymbolen (vorerst nur Konsonanten). Der Grad der Abweichung von Segmentgrenzen für bestimmte Laute und Lautklassen, der bei menschlichen Segmentierern zu beobachten ist, wird als erlaubte Abweichung der automatischen Segmentation von den manuellen Referenzsegmentationen festgelegt. Unser Vorgehen bei der Evaluation war also wie folgt:

A Vergleich der Segmentationen menschlicher Segmentierer untereinander in einem One-Leave-Out-Verfahren in Bezug auf
- gewählte Transkriptionssymbole (Konsonanten)
- gewählte Segmentgrenzen bei übereinstimmenden Transkriptionssymbolen

B Vergleich der automatischen Segmentation mit manuellen Segmentationen in einem One-Leave-Out Verfahren in Bezug auf
- gewählte Transkriptionssymbole (Konsonanten)
- gewählte Segmentgrenzen bei übereinstimmenden Transkriptionssymbolen

C Gegenüberstellung der Daten von A und B

4.2 Ergebnisse

4.2.1 Sprachmaterial

Das anaysierte Sprachmaterial stammt aus dem PhonDat-II Zugauskunft-Korpus mit gelesenen Sprachdaten von 10 Sprechern. Neben der vollständigen automatischen Segmentation des Materials eines Sprechers sind jeweils 64 Sätze von mindestens einem bis zu drei menschlichen Segmentierern bearbeitet worden.

4.2.2 Übereinstimmung der Label

Tabelle 2 gibt eine Übersicht über die prozentuale Übereinstimmung der transkribierten Konsonanten. In einem ersten Block sind die Zahlen für die menschlichen Segmentierer aufgelistet, im zweiten Block die Zahlen für den Vergleich der automatischen Segmentation mit den manuellen. Es sind Daten für einzelne Konsonanten, für Konsonantenklassen und für die Konsonanten im Ganzen errechnet worden.

Es zeigt sich, daß die Übereinstimmung bei der Wahl des geeigneten Symbols bei Plosiven sowohl bei den menschlichen Segmentierern (89,9%) als auch bei der automatischen Segmentation im Vergleich zur Referenz (80,2%) am niedrigsten ist. Bei den anderen Lautklassen ist die Übereinstimmung jeweils größer mit 98,0% bei Frikativen, 97,5% bei Nasalen, 98,0% bei /l/ und 96,0 % bei /r/ unter menschlichen Segmentierern und 80,2% bei Frikativen, 94,4% bei Nasalen, 94,4% bei /l/ und

99,0% bei /r/ im Vergleich der automatischen Segmentation mit den Referenzsegmentationen. Insgesamt besteht eine Übereinstimmung von 94,8% in der Wahl der Konsonantenbezeichnung unter den menschlichen Segmentierern gegenüber 88,4% bei der automatischen Segmentation im Vergleich zu den Referenzsegmentationen.

		Vergleich von manuellen Segmentationen untereinander			Vergleich von manuellen Segmentationen vs. MAUS-Segmentationen		
		gemittelte man. Seg. = Referenz	Übereinstimmungen d. Segment. untereinander		manuelle Seg. = Referenz	Übereinstimmungen von MAUS	
		Anzahl = 100%	Anzahl	in %	Anzahl = 100%	Anzahl	in %
Plosive	p	384	360	93,75	212	162	76,41
	b	2593	2534	97,72	1594	1315	82,49
	t	6236	5771	92,54	3788	3039	80,22
	d	2390	1902	79,58	1343	1008	75,05
	k	3156	2906	92,07	1928	1719	89,15
	g	2012	1728	85,88	1176	848	72,1
	Q	3906	3384	86,63	2357	1846	78,31
	gesamt	20677	18585	89,88	12398	9937	80,15
Frikative	f	3456	3428	99,18	2155	2146	99,6
	v	1216	1174	96,54	730	644	88,2
	s	4001	3939	98,45	2440	2322	95,1
	z	560	520	92,85	362	357	98,61
	S	633	628	99,21	368	346	94,02
	C	2662	2618	98,34	1625	1532	94,27
	j	330	318	96,36	202	197	97,52
	x	2307	2294	99,43	1432	1330	92,87
	h	1116	1030	92,29	691	494	71,49
	gesamt	16281	15949	97,96	10005	9368	93,63
Nasale	m	4562	4480	98,2	2829	2743	96,96
	n	10710	10484	97,88	6552	6220	94,93
	N	1680	1570	93,45	1027	858	83,54
	gesamt	16952	16536	97,53	10408	9821	94,36
Lateral	l	1527	1496	97,96	872	559	64,1
Vibrant	r	1519	1458	95,98	931	922	99,03
gesamt		**56056**	**54022**	**94,84**	**34614**	**30607**	**88,42**

Tabelle 2: Übereinstimmung bei gewählten Transkriptionssymbolen für Konsonanten

4.2.3 Übereinstimmung der Segmentgrenzen

Die folgende Tabelle 3 zeigt die mittlere Abweichung von gewählten Segmentgrenzen bei übereinstimmenden Transkriptionssymbolen. Eine Segmentgrenze wird dabei nicht in Bezug auf einen bestimmten Laut beschrieben, sondern in Bezug auf zwei aneinandergrenzende Laute bzw. Äußerungsbeginn und -ende.

Man kann sehen, daß MAUS besonders dann große Abweichungen bei der Festlegung einer Segmentgrenze zeigt, wenn es um die Bestimmung des Äußerungsbeginns bzw. -endes geht. Dieses Problem kann mit relativ wenig Aufwand im Refinement behandelt und damit prozentual eine große Verbesserung der Gesamtsegmentationsleistung erreicht werden.

Besondere Schwierigkeiten machen u.a. Nasal- und Lateralübergänge, die auch für den Menschen nicht immer einfach zu bestimmen sind.

Als ausreichend gut können unabhängig von einem Vergleich mit den Segmentaitonsleistungen menschlicher Segmentierer Abweichungen bis zu 10 ms gewertet werden. Akzeptabel sind Abwei-

chungen bis zu 20 ms, jedoch nicht bei allen Lautklassen (etwa bei Plosiven). Am dringendsten sind Verbesserungen bei Lautübergängen mit Abweichungen über 20 ms, allerdings auch in Abhängigkeit von der Lautklasse.

Lautübergang	MAUS mittlere Abweichung in ms	Mensch mittlere Abweichung in ms
BEG -> sthFrik	454	5
BEG -> sthPlos	391	11
BEG -> stlPlos	342	11
BEG -> Nasal	195	2
Nasal -> END	165	10
stlPlos -> sthFrik	116	15
Lateral -> Nasal	75	3
stlPlos -> sthPlos	75	22
Lateral -> END	71	18
stlPlos -> stlPlos	61	10
Vokal -> END	61	10
stlPlos -> END	47	27
Lateral -> sthPlos	45	3
Nasal -> Nasal	43	10
stlFrik -> END	38	19
Vokal -> Lateral	36	8
Nasal -> sthFrik	33	5
Vokal -> sthFrik	29	3
stlosFrik -> sthFrik	27	10
stlPlos -> Vokal	23	1
stlFrik -> stlPlos	21	3
Vokal -> Vokal	20	9
stlFrik -> sthPlos	19	4
Nasal -> Lateral	19	5
Vokal -> Nasal	19	2
Vokal -> sthPlos	19	3
Nasal -> Vokal	18	1
Lateral -> Vokal	17	2
Nasal -> sthPlos	15	5
sthPlos -> R	15	3
stlFrik -> Nasal	14	5
Vokal -> R	14	5
stlFrik -> stlFrik	13	10
R -> Vokal	13	7
stlFrik -> R	12	8
sthFrik -> Vokal	12	6
sthPlos -> Vokal	12	1
Lateral -> stlFrik	11	6
Nasal -> R	11	2
Vokal -> stlPlos	11	7
Nasal -> stlPlos	10	5
stlPlos -> stlFrik	10	6
stlPlos -> Lateral	9	7
sthPlos -> Nasal	9	3
Lateral -> stlPlos	8	7
Vokal -> stlFrik	8	3
Nasal -> stlFrik	7	3
stlPlos -> Nasal	7	4
stlFrik -> Vokal	6	2
sthPlos -> Lateral	6	2

Tabelle 3: Abweichung bei der Wahl der Segmentgrenzen

Eine andere Art der Segmentationsleistung wird durch die Anzahl der Segmentgrenzen, die inner-

halb eines bestimmten Zeitbereichs von den Referenzgrenzen abweichen, bestimmt. Tabelle 4 zeigt diese Angaben für die automatische Segmentation im Vergleich mit den menschlichen Segmentationen.

Bereich	MAUS	Menschen
= 0 ms	1 %	63 %
< 5 ms	36 %	73 %
< 10 ms	61 %	87 %
< 15 ms	76 %	93 %
< 20 ms	84 %	96 %
< 32 ms	90 %	99 %
< 64 ms	95 %	100 %

Tabelle 4: übereinstimmende Segmentgrenzen innerhalb bestimmter Zeitbereiche

5 Zusammenfassung

Im vorliegenden Dokument wurde das Münchener AUtomatische Segmentationssystem - MAUS - vorgestellt, das im Verbmobil-Teilprojekt 14.7 entwickelt wird. Mit MAUS ist es möglich, das akustische Signal einer sprachlichen Äußerung (deutsch), für das eine orthographische Repräsentation vorliegt, automatisch zu transkribieren und segmentieren. Das System basiert auf einer statistischen Modellierung unter Einbeziehung phonetischen Wissens über segmentale Prozesse gesprochener Sprache.

Eine gründliche automatische Evaluation der Analyseleistung von MAUS wurde durchgeführt, wobeiautomatische Segmentationen manuell erstellen Referenzsegmentationen gegenübergestellt wurden.

Die Evaluation bezieht sich in einem ersten Teil auf die Übereinstimmung der automatischen mit den manuellen Segmentationen bei der Wahl der Transkriptionssymbole. Dabei zeigte sich eine Übereinstimmung von 88,3% gegenüber einer Übereinstimmung von 94,8%, die innerhalb einer Gruppe menschlicher Segmentierer erreicht wird.

In einem zweiten Teil wurde die Übereinstimmung bei der Bestimmung der Grenzen gleich etikettierter Segmente betrachtet. Durch die Bezugnahme auf Lautübergänge erhält man Anhaltspunkte für die präzise Beschreibung der Schwächen des Systems, welche in einer der statistischen Analyse nachgeschalteten Refinement-Stufe mit signalnahen Methoden gezielt behandelt werden. Besonders große Abweichungen zeigen sich am Beginn und Ende einer Äußerung, aber z.B. auch an Grenzen mit beteiligten Nasalen oder Lateralen.

In einem Bereich bis zu 10 ms stimmen 61% der automatisch bestimmten Grenzen überein (menschliche Segmentierer: 87%), in einem Bereich bis zu 20 ms sogar 84% (menschl. Segmentierer: 96%) .

Dies ist ein beachtliches Ergebnis, das durch Ausbau der Refinement-Stufe weiter verbessert werden kann. In Vorbereitung ist die Software-Anpassung zur automatischen Segmentation von Verbmobil-Sprachaufnahmen.

6 Literatur

[1] Eisen, B., H. G. Tillmann, Chr. Draxler (1993): “Consistency of Judgements in Manual Labelling of Phonetic Segments: The Distinction between Clear and Unclear Cases”, in: *Proceedings of ICSLP 1993*, Banff - Canada, pp.871 - 874.

[2] Huang, X.D., Jack, M. A. (1988): “Hidden Markov Modelling of Speech Based on Semicontinuous Models”, in: Electronic Letters, Vol. 24, No.2, 7. Januar 1988, pp. 6-7.

[3] Wesenick, M.-B.(1994): Entwurf eines Regelsystems der Aussprache des Deutschen als Basis für empirische Untersuchungen. Magisterarbeit, IPSK, Universität München 1994.

[4] Wesenick, M.-B., Schiel, F. (1994): “Applying Speech Verification to a Large Data Base of German to Obtain a Statistical Survey about Rules of Pronunciation”, in: *Proceedings of ICSLP 1994*, Yokohama - Japan, pp. 279 - 282.

[5] Wesenick, M.-B. (1995): ”Regelsystem zur Generierung von Aussprachevarianten - Bestandteil des Münchener Automatischen Segmentationssystems (MAUS)”, Verbmobil-Memo Nr. 96.