Das Projekt GlobalPhone: Multilinguale Spracherkennung

Tanja Schultz und Alex Waibel

Zusammenfassung

Dieser Artikel beschreibt die Entwicklung eines Spracherkennungssystems GlobalPhone für den Einsatz in multilingualer kontinuierlicher Spracherkennung für große Vokabulare. In diesem Projekt der Universität Karlsruhe werden Systeme für 15 Sprachen untersucht: Arabisch, Chinesisch (Mandarin und Wu Dialekt), Deutsch, Englisch, Französisch, Japanisch, Koreanisch, Kroatisch, Portugiesisch, Russisch, Spanisch, Schwedisch, Tamil, und Türkisch. Anhand fünf dieser Sprachen entwickelten wir ein universelles Phonemset, auf dessen Basis die sprachenspezifischen akustischen Modelle zu einem multilingualen Modul kombiniert werden. Kontextabhängige multilinguale Modelle werden dabei durch Integration von Kontextfragen nach Sprachen und Sprachgruppen erzeugt und die Resultate analysiert.

This paper describes our recent effort in developing the GlobalPhone recognizer engine for multilingual large vocabulary continuous speech recognition. This project at the University of Karlsruhe investigates LVCSR systems in 15 languages, namely Arabic, Chinese (Mandarin and Wu dialect), Croatian, English, French, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. We present results of our experiments towards multilinguality using the described database. We developed a global phoneme set and built a multilingual speech recognition system by combining acoustic models of five different languages. Context dependent phoneme models are created using questions about languages and language groups.

1 Einleitung

Angesichts der zunehmenden Kooperation vieler Menschen unterschiedlichster Sprachen erscheinen multilinguale sprachverarbeitende Systeme immer dringlicher. Je nach Definition existieren auf der Welt zwischen 3000 und 6000 Sprachen, davon werden ca. 200 von mehr als einer Millionen Menschen gesprochen. Das Projekt GlobalPhone der Universität Karlsruhe beschäftigt sich mit der Entwicklung multilingualer sprachverarbeitender Systeme unter den folgenden beiden Gesichtspunkten:

1. Die Entwicklungszeiten für Spracherkennungssysteme in neuen Sprachen sind sehr lang; die Sammlung der dazu bisher notwendigen Sprachdaten zu kostspielig. Im Rahmen des Projektes wird nach Lösungen

gesucht, die eine Verkürzung des Entwicklungszyklus erlauben und die Menge der dazu notwendigen Sprachdaten reduzieren.

2. Für Übersetzungsaufgaben in multilingualen Dialogsystemen sind mehrere Eingabesprachen zu bewältigen. Unter GlobalPhone soll ein System entwickelt werden, das einen Wechsel der Eingabesprache im laufenden Betrieb toleriert und gegebenenfalls notwendige sprachenspezifische Komponenten eigenständig nachlädt. Dazu muß die Eingabesprache zuvor zuverlässig identifiziert werden, was auch für multilinguale Informationssysteme notwendig ist.

In GlobalPhone wird ein multilinguales Spracherkennersystem realisiert, das die akustischen Modelle der beteiligten Sprachen kombiniert. Dazu muß zunächst ein multilinguales oder *universelles* Phonemset entwickelt werden, das einen möglichst breiten Sprachraum abdeckt.

2 Die GlobalPhone Datenbasis

Für diese Forschung ist eine multilinguale Sprachdatenbasis erforderlich, mit deren Sammlung im Projekt GlobalPhone Mitte 1996 begonnen wurde. Mittlerweile wurden Daten von über 1200 Sprechern 13 verschiedener Sprachen gesammelt.

2.1 Korpusdesign

Zum Trainieren und Testen stabiler akustischer Phonemmodelle sollten pro Sprache etwa 10.000 transkribierte Äußerungen einer repräsentativen Auswahl von Muttersprachlern verfügbar sein. Die wesentlichen Kriterien für eine Sprachdatensammlung sind die Faktoren Kosten und Zeit. Um die zeitaufwendige Verschriftung von Sprache zu vermeiden, wurden daher für die GlobalPhone Datenbasis ausschließlich gelesene Sprachdaten von überregionalen, im Internet vertretenen Tageszeitungen des jeweiligen Landes gesammelt. Um den Gebrauch sprachübergreifender Begriffe und Eigennamen zu gewährleisten, wählten wir Berichte über das politisches Geschehen und Wirtschaftsthemen. Dieses Vorgehen resultiert in sehr großen Wortschätzen und gewährleistet gleichzeitig die Vergleichbarkeit zwischen den Sprachen. Die elektronische Verfügbarkeit auf dem Internet ermöglicht das Sammeln zusätzlicher großer Textkorpora zur zuverlässigen Schätzung der Sprachmodelle.

2.2 Sprachenauswahl

	Sprache	Sprachraum	Sprecher	Sprachfamilie
1	Mandarin*	China	907 Mio	Sino-Tibetan (Sinitisch)
2	Englisch	USA, UK, Can, Australien	456 Mio	I-E (Germanisch)
3	Hindi	Indien	383 Mio	I-E (Indo-Iranisch)
4	Spanisch*	Latein-Amerika, Spanien	362 Mio	I-E (Romanisch)
5	Russisch*	Rußland, unabh. Staaten	293 Mio	I-E (Slawisch)
6	Arabisch*	Nord-Afrika, mittl. Osten	208 Mio	Afro-Asiatisch (Semit.)
7	Bengalisch	Bangladesch, Indien	189 Mio	I-E (Indo-Iranisch)
8	Portugiesisch*	Brasilien, Portugal, Angola	177 Mio	I-E (Romanisch)
9	Malay-Indo.	Indien, Malaysia, Brunei	148 Mio	Austronesisch (Polyn.)
10	Japanisch*	Japan	126 Mio	isolierte Sprache
11	Französisch	F, Can, Afrika, Schweiz	123 Mio	I-E (Romanisch)
12	Deutsch*	D, Österreich, Schweiz	119 Mio	I-E (Germanisch)
15	Koreanisch*	Korea, China	73 Mio	isolierte Sprache
17	Tamil*	Indien, Sri Lanka, Malaysia	67 Mio	Dravidisch
20	Wu*	China (Shanghai)	64 Mio	Sino-Tibetan (Sinitisch)
21	Italienisch	Italien, Schweiz	63 Mio	I-E (Romanisch)
25	Türkisch*	Türkei	57 Mio	Altaisch (Turksprache)
43	Kroatisch*	Balkan	20 Mio	I-E (Slawisch)
85	Schwedisch*	Schweden, Finnland	9 Mio	I-E (Germanisch)

Tabelle 1: Die wichtigsten Weltsprachen nach [1] (I-E=Indo-Europäisch)

Tabelle 1 vermittelt einen Überblick über die Sprecherzahlen und den Sprachraum der häufigsten Sprachen der Welt [1]. Die Sprachen der GlobalPhone Datenbasis wurde nach Verbreitungsgrad, wirtschaftlicher Bedeutung, Zukunftspotential, sowie experimentellen Erwägungen ausgewählt. Momentan besteht die Datenbasis aus den in Tabelle 1 mit '*' markierten Sprachen. Berücksichtigt man die Tatsache, daß in Englisch mit „Wall Street Journal“ und in Französisch mit „Bref“ bereits domänegleiche Datenbasen verfügbar sind, können damit 9 der 12 wichtigsten Weltsprachen abgedeckt werden.

2.3 Datensammlung und -aufbereitung

Eine Sammlung der Daten via Telefon (Call Home, OGI) wurde nicht in Betracht gezogen, da durch die eingeschränkte Bandbreite (350 - 3500 Hz) wichtige Besonderheiten einer Sprache möglicherweise nicht erfaßt werden könnten. Auf Sprachdaten von Sprechern, die nicht in ihrem Heimatland leben, wurde ebenfalls verzichtet, da dies die Muttersprache verfälschen kann. Die Sammlung wurde daher direkt im Mutterland der jeweiligen Sprecher vorgenommen. Jeder Sprecher erhielt die Aufgabe etwa 20 Minuten Zeitungstexte vorzulesen. Die Sprache wurde mit einem tragbaren DAT-Rekorder Sony TDC-8 und einem Nahsprechmikrofon Sennheiser HD-440-6 aufgezeichnet. Die Transkriptionen wurden anschließend von Muttersprachlern validiert, wobei Markierungen für artikulatorische Geräusche hinzugefügt wurden. Tabelle 2 zeigt den aktuellen Stand der GlobalPhone Datenbasis. Die in Klammern zugefügten Sprecherzahlen zeigen an, daß in diesen Sprachen die Datenbasis derzeit erweitert wird.

Sprache	Sprecher Anzahl	Äußerungen Anzahl	Audio Std	lfd. Worte (in Tsd.)	Vokabular (in Tsd.)
Arabisch	100	i.B.	20	i.B.	i.B.
Japanisch	121	9785	25	204K	23K
Koreanisch	100	6868	18	80K	40K
Kroatisch	(+30) 83	4019	14	106K	20K
Mandarin	132	9103	28	250K	12K
Portugiesisch	(+50) 74	6726	17	126K	6K
Russisch	(+50)100	10229	20	155K	22K
Schwedisch	100	i.B.	20	i.B.	i.B.
Spanisch	100	6866	22	176K	21K
Tamil	49	i.B.	12	i.B.	i.B.
Türkisch	100	6872	17	112K	16K
Wu	40	3000	10	80K	8K
Deutsch	(+80) 19	3300	10	47K	10K

Tabelle 2: Die GlobalPhone Datenbasis (i.B.=in Bearbeitung)

3 Multilinguale akustische Modellierung

Ein wesentlicher Aspekt der multilingualen Spracherkennung ist die akustische Modellierung von Gemeinsamkeiten verschiedener Sprachen. So können die akustischen Modelle derjenigen Phoneme kombiniert werden, die in mehr als einer Sprache vorkommen. Dadurch erzielt man mehrere Vorteile: erstens reduziert sich die Anzahl zu modellierender Parameter und damit die Komplexität des Gesamtsystems und zweitens erhält man eine günstige Ausgangsbasis für die Modellierung neuer, noch nicht initialisierter Sprachen. Diejenigen Phoneme, die nur in einer Sprache vorkommen, können zur Sprachenidentifizierung eingesetzt werden.

Phoneme [Worldbet]	KO	SP	CR	TU	JA	$\sum$
n,m,s,l,tS,p,b,t,d,g,k,	X	X	X	X	X	
i,e,o	X	X	X	X	X	14
f,j,z		X	X	X	X	
r,u	X	X	X	X		
dZ	X		X	X	X	6
a	X	X	X			
S			X	X	X	
h	X			X	X	
4	X	X			X	4
$\tilde{n}$,x,L		X	X			
A				X	X	
N	X	X				
V,Z			X	X		
y,7	X			X		
ts			X		X	10
p',t',k',dZ',s',oE,oa,4i,uE	X					
E,$\wedge$,i$\wedge$,u$\wedge$,iu,ie,io,ia	X					17
D,G,T,V,r(,		X				
ai,au,ei,eu,oi,a+,e+,i+,o+,u+		X				15
palatal c, palatal d			X			2
ix, weichzeichen				X		2
?,Nq,V[,A:,e:,i:,o:,4:					X	8
Monolingual $\sum = 180$	40	40	30	29	31	
Multilingual						78

Tabelle 3: IPA-basierte Phonemklassen [Worldbet Notation]

Für die multilinguale akustische Modellierung muß die Ähnlichkeit zwischen zwei Phonemen bestimmt werden. Diese könnte mit Hilfe bereits existierender phonetischer Inventare wie Sampa, Worldbet oder IPA apriori bestimmt werden. Es wurden auch bereits datengetriebene Algorithmen zum Auffinden von Ähnlichkeiten vorgeschlagen [2], [3]. Sie berücksichtigen jedoch nicht die Modellierung kontextabhängiger Polyphone zur Spracherkennung großer Vokabulare. Wir schlagen daher einen datengetriebenen Ansatz vor, der akustische Ähnlichkeiten zwischen Sprachen ausnutzt und gleichzeitig die kontextabhängige Modellierung von multilingualen Phonemen ermöglicht. Die Experimente wurden mit den fünf Sprachen Japanisch, Kroatisch, Koreanisch, Spanisch und Türkisch aus der GlobalPhone Datenbasis durchgeführt.

3.1 Globales Phonemset

Ausgangspunkt unserer Experimente ist eine Einteilung der sprachenspezifischen Phoneme in Phonemklassen gemäß des IPA Schemas [4]. Phoneme verschiedener Sprachen, die durch dasselbe IPA Symbol repräsentiert sind, werden dabei zunächst zu einer gemeinsamen Klasse zusammengefaßt. Das so entstehende Phonemset ist in Tabelle 3 in Worldbet Notation dargestellt. Das gesamte Set für die fünf Sprachen besteht aus 78 Phonemen zuzüglich eines Modells für Stille und 2 Modellen für artikulatorische Geräusche.

Etwa die Hälfte aller Phoneme werden von mehreren Sprachen geteilt, allein 14 Phoneme kommen in allen 5 Sprachen vor. Abbildung 1 zeigt am Beispiel der häufigsten Konsonanten, daß die relative Auftrittshäufigkeit im GlobalPhone-Korpus sehr stark von der Sprache abhängt.

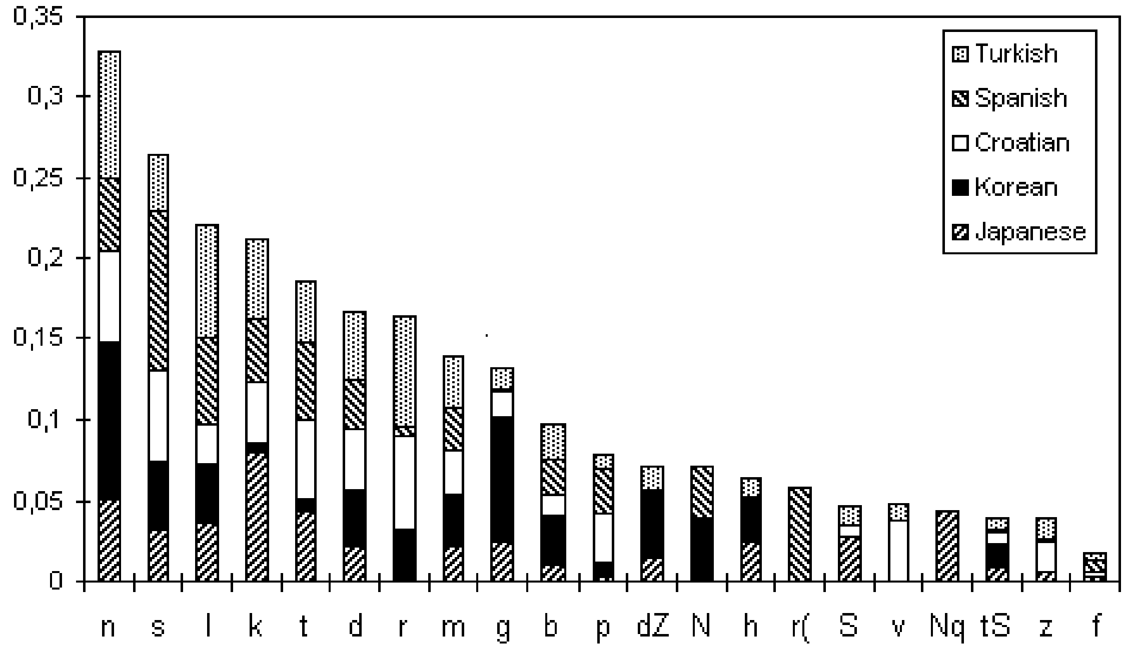

Abbildung 1: Relative Häufigkeit von Konsonanten in fünf Sprachen

3.2 Multilinguale Kontextmodellierung

In den heute gängigen Erkennersystemen werden Phonemmodelle zu kontextabhängigen Modellen geballt, um einen möglichst guten Arbeitspunkt zwischen Generalisierungsfähigkeit und Modellgenauigkeit zu erzielen. In unserem Erkenner verwenden wir generalisierte Subpolyphone, wobei der

Kontext der zu generalisierenden Untereinheiten bis in die angrenzenden Worte reichen kann. Wir benutzen ein entscheidungsbaumbasiertes divisives Ballungsverfahren mit einem Entropie-Distanzmaß, das durch die Distanz der Entropien der Mixturgewichteverteilungen der einzelnen Modelle definiert ist. Es werden nur gleichartige Zustände (wir unterscheiden einen Beginn-, einen Mitte- und einen Endzustand) desselben Phonems geballt. Dazu wird ein Fragenkatalog erstellt, der aus phonetisch motivierten Fragen über den Kontext von Phonemen besteht. Aus diesem Katalog wird jeweils diejenige Frage ausgewählt, bei deren Anwendung der Entropieverlust durch Aufteilen des Knotens in zwei Nachfolgerknoten am größten wird. Der Ballungsvorgang endet, wenn eine vordefinierte Anzahl von Ballungsknoten erreicht ist.

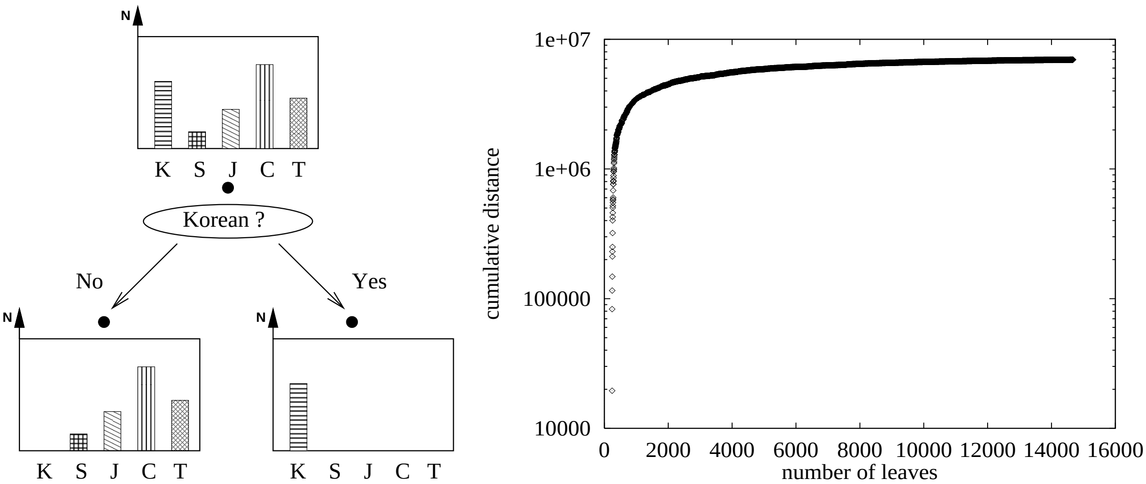

Abbildung 2: Analyse der Sprachenfragen

Wir erweitern dieses Verfahren auf den multilingualen Fall, indem wir dem bisherigen Fragenkatalog Fragen nach der Sprache eines Phonemes zufügen. Beim Ballungsvorgang entscheiden nun die Daten darüber, ob Fragen nach der Sprache bedeutungsvoller sind als Fragen nach dem phonetischen Kontext. Durch die Analyse des entstandenen Entscheidungsbaumes wollten wir untersuchen, wie schnell die Spracheninformation beim Ballungsvorgang ausdifferenziert wird. Falls sich Phoneme über Sprachen hinweg stark unterscheiden, wird das Ballungsverfahren diese Modelle früh aufspalten und letztlich mit monolingualen Modellen enden. Falls die Information über die Sprache jedoch nicht relevant ist, entstehen beim Ballen „echte“ multilinguale Modelle. Wir berechneten daher für jeden Ballungsknoten das Maß der Entropiedistanz $H_D = H_{org} - (H_{yes} + H_{no})$ zwischen der Entropie des Knotens vor seiner Aufspaltung H_{org} und den beiden Nachfolgerknoten nach der Aufspaltung (H_{yes}, H_{no}). Die Entropie errechnet sich dabei aus der Verteilung über die Sprachen im jeweiligen Knoten, wie im linken Teil der Abbildung 2 veranschaulicht. In der rechten Graphik von Abbildung 2 ist das summierte Distanzmaß $\sum H_D$ über alle Ballungsknoten aufgetragen. Man sieht, daß der größte Anteil an Spracheninformation nach etwa 2000 Aufspaltungen ausdifferenziert ist. Unser multilinguales Erkennersystem, das bei 5 Sprachen mehr als 2000 Polyphone modelliert, besteht demnach größtenteils aus monolingualen Modellen.

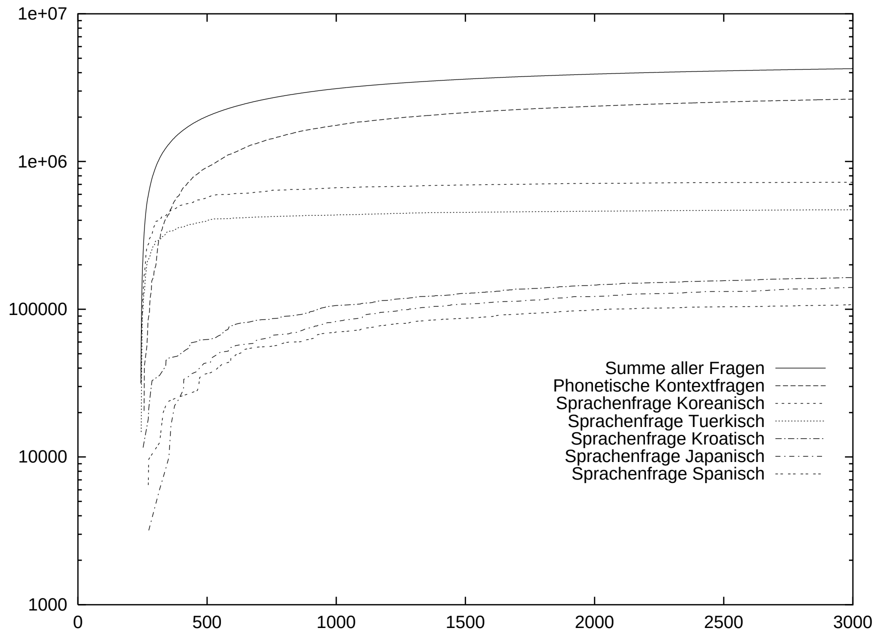

Abbildung 3: Sprachenspezifischer summierter Entropieverlust

Anschließend untersuchten wir, welchen Stellenwert die Sprachenfragen im Verhältnis zu phonetischen Kontextfragen einnehmen. Dazu verglichen wir den summierten Entropieverlust, der aus Kontextfragen resultiert, mit dem Verlust, der sich aus den Sprachenfragen ergibt. Die Kurve *Summe aller Fragen* in Abbildung 3 gibt den summierten Entropieverlust für alle gestellten Fragen über die ersten 3000 Ballungsknoten wieder; die Kurve *Phonetische Kontextfragen* zeigt den Verlauf für solche Fragen, die sich nicht auf Sprachen beziehen. Die Differenz verdeutlicht, daß ein beträchtlicher Anteil des Entropieverlustes durch Sprachenfragen entsteht. Dies deckt sich mit den Resultaten aus Abbildung 2. Die übrigen fünf Kurven zeigen, in welchem Maß eine einzelne Sprache am Entropieverlust beteiligt ist: Koreanisch und Türkisch tragen zu einem hohen Verlust bei. Diese Sprachenfragen werden viel früher gestellt als Kroatisch, Japanisch oder Spanisch, was nahelegt, daß sich die akustischen Modelle der ersten beiden Sprachen stark von denen der letzten drei Sprachen unterscheiden.

#	500 Modelle	#	1000 Modelle	#	1500 Modelle	#	3000 Modelle
76	KO+TU	92	KO+TU	100	KO+TU	146	word bound
38	KOREAN	54	KOREAN	73	KOREAN	131	back-vow
30	front-vow	48	back-vow	73	back-vow	130	front-vow
27	back-vow	45	front-vow	65	front-vow	128	consonant
23	vowel	38	unvoiced	61	word bound	113	KO+TU
22	unvoiced	37	word bound	53	consonant	98	KOREAN
20	silence	36	vowel	48	unvoiced	97	voiced
19	fric-sibil	32	consonant	48	alveodental	90	vowel
16	word bound	29	silence	46	vowel	88	unvoiced
14	nasal	28	voiced	42	voiced	85	nasal
10	voiced	26	nasal	42	nasal	84	alveodental
10	round	25	frik-sibil	36	silence	79	JAPANESE
10	JAPANESE	24	plos-unvoic	36	plos-unvoic	63	plos-unvoic
10	consonant	23	alveodental	35	frik-sibil	59	frik-sibil
9	plos-unvoic	22	round	32	JAPANESE	59	close-vow
9	open-vow	19	plosive	29	round	56	silence
9	CR+JA+SP	19	JAPANESE	28	plosive	55	round
8	vow-a	16	open-vow	24	CR+SP	54	plosive
8	plosive	16	CR+JA+SP	23	open-vow	47	CROATIAN

Tabelle 4: Häufigkeiten von Kontextfragen beim divisiven Ballen

Im letzten Experiment faßten wir Sprachen zu verschiedenen Gruppen zusammen und fügten die Fragen nach diesen Sprachengruppen in den Fragekatalog ein. Die Häufigkeit, mit der die Sprachgruppenfragen während des Ballungsvorgangs gestellt wurden, analysierten wir an vier Abbruchstellen: nach jeweils 500, 1000, 1500 und 3000 geballten Polyphonmodellen. Die Tabelle 4 zeigt die nach Häufigkeit sortierten Fragen für die 500, 1000, 1500 und 3000 Modelle. Wie nach den bisherigen Ergebnissen wenig verwundert, dominiert die Sprachgruppe Koreanisch+Türkisch (KO+TU). Erstaunlich ist allerdings, wie häufig diese Frage gerade zu Beginn des Ballungsverfahrens gestellt wird. Die nächst wichtigste Frage ist die nach der Sprache Koreanisch, erst mit einigem Abstand folgt dann die nach Japanisch.

Nach diesen Analysen läßt sich zusammenfassend sagen, daß:

- das angewendete Ballungsverfahren sehr schnell die Information „Sprache“ ausdifferenziert und dadurch eher monolinguale kontextabhängige Modelle entstehen
- koreanische und türkische Modelle sich von den restlichen stark unterscheiden.

System Parameter	Monolingual 5 x 1500	Multilingual 7500 Modelle	Multilingual 3000 Modelle
Japanisch	13.0	-	15.0
Koreanisch	47.3	47.7	49.0
Kroatisch	26.9	30.2	31.9
Spanisch	27.6	30.0	32.4
Türkisch	20.1	21.3	21.3

Tabelle 5: Erkennungsleistungen für fünf Sprachen [Wortfehlerrate]

4 Entwicklung eines multilingualen Spracherkennungssystems

Mit dem beschriebenen Ballungsverfahren entwickelten wir ein multilinguales Spracherkennungssystem für große Vokabulare, das die akustischen Modelle der fünf Sprachen Japanisch, Koreanisch, Kroatisch, Spanisch und Türkisch in ein gemeinsames Modul kombiniert. Tabelle 5 vergleicht die Wortfehlerrate des multilingualen Systems mit den fünf monolingualen. Obwohl die Anzahl der Parameter im multilingualen System um 40% geringer ist (3000 Modelle) als in den monolingualen Systemen (5x1500 Modelle), verliert das System dabei im Mittel noch akzeptable 2.9% (1.2% - 5%) Wortakkuratheit. Um zu überprüfen, ob der gesamte Verlust durch die geringere Anzahl der Modelle erklärt werden kann, bildeten wir auch ein multilinguales System mit 7500 Modellen. Der Vergleich der Wortakkuratheiten zeigt, daß dennoch ein mittlerer Verlust von 1.6% (0.4% - 3.3%) bestehen bleibt. Diese Differenz weist darauf hin, daß die multilingualen Phoneme an Modellgenauigkeit einbüßen.

5 Ausblick

Ziel unserer zukünftigen Arbeiten wird zunächst die Verbesserung der monolingualen Systeme sein, um einen negative Einfluß unscharfer Modelle auf die multilingualen Modelle auszuschließen. Ein weiterer Arbeitsschwerpunkt liegt im Bereich der Initialisierung neuer, noch nicht modellierter Sprachen.

6 Danksagung

Die Autoren danken den Mitgliedern des GlobalPhone Teams, ohne deren Enthusiasmus beim Sammeln und Validieren der Datenbasis diese Forschungsarbeit nicht möglich gewesen wäre: Olfa Karboul-Zouari und Mohamed Zouari (Arabisch), Jürgen Reichert, Jing Wang, Tianshi Wei und Jiaxing Weng (Chinesisch), Hiroko Akatsu, Laura J. Mayfield-Tomokyo und Sayoko Takeda (Japanisch), Keal-Chun Cho, Daniel Kiecza und Sang-Hun Shin (Koreanisch), Sanela Habibija und Stefan Raschke (Kroatisch), Orest und Natalia Mikhailiouk (Russisch), Raul Ivo Faller und Caleb Everett (Portugiesisch), Giovanni Najera Barquero (Spanisch), sowie Kenan Çarki und Mutlu Yalcin (Türkisch).

Literatur

[1] *Webster's New Encyclopedic Dictionary.* Black Dog & Leventhal 1992.

[2] O. Andersen, P. Dalsgaard, and W. Barry: *Data-Driven identification of Poly- and Mono-phonemes for four European Languages* in: Proc. Eurospeech, pp. 759-762, Berlin 1993.

[3] J. Köhler: *Multi-lingual Phoneme Recognition exploiting Acoustic-phonetic Similarities of Sounds* in: Proc. ICSLP, pp. 2195-2198, Philadelphia 1996.

[4] The IPA 1989 Kiel Convention. In: Journal of the International Phonetic Association 1989(19) pp. 67-82

[5] T. Schultz and A. Waibel: *Multilingual and Crosslingual Speech Recognition* in: Proceedings of the DARPA Broadcast News Workshop 1998.