Fraunhofer
IKTS

FRAUNHOFER INSTITUTE FÜR KERAMISCHE TECHNOLOGIEN UND SYSTEME IKTS

MACHBARKEITSSTUDIE

Zur automatischen Spracherkennung des Obersorbischen

Ivan Kraljevski
Marek Rjelka
Frank Duckhorn
Constanze Tschöpe
Christian Richter*
Matthias Wolff*

Fraunhofer Institut für Keramische Technologien und Systeme IKTS
Maria-Reiche-Straße 2, 01109 Dresden

**Brandenburgische Technische Universität Cottbus–Senftenberg, Cottbus*

Kunde: Stiftung für das sorbische Volk, Postplatz 2, 02625 Bautzen
Angebotsnummer: 068-A20-0088-346
Auftrag vom: 28.07.2020
Referenznummer: 068-200025
Berichtnummer: 01

Vetraulichkeit: frei

Dresden, Dezember 2020

Externe Veröffentlichung

Abteilungsleiter

Projektleiter

Inhalt

1 Einleitung

In diesem Bericht werden die Ergebnisse der Machbarkeitsstudie zur automatischen Erkennung der obersorbischen Sprache dargestellt.

Hauptziel dieser Untersuchung war es, bestehende Akustikmodelle mittels Transferlernen an das Obersorbische (ISO-Sprachencode: hsb) als Beispiel einer gefährdeten Sprache anzupassen und die resultierenden Modelle zu evaluieren. Dazu wurde ein prototypisches automatisches Spracherkennungssystem (ASR) entwickelt, das als Demonstrator auf einem eingeschränkten Sprachbereich fungiert.

Der Bericht ist einzelne, den Arbeitspaketen entsprechende, Abschnitte unterteilt. Jeder dieser Abschnitte beginnt mit einer kurzen Vorstellung der Ziele, gefolgt von einer detaillierten Beschreibung der Aktivitäten. Alle involvierten Personen trugen zum erfolgreichen Abschluss dieser Arbeit bei.

2 AP 1: Graphem- und Phonembestand

Die Voraussetzung für dieses Arbeitspaket war einen präzise Definition des Graphem- und Phonembestands.

Die Quellen dieser Informationen waren die Wikipedia-Seite der obersorbischen Sprache (https://en.wikipedia.org/wiki/Upper_Sorbian_language), die Seite obersorbisch.de des Sorbischen Instituts und ein Exzerpt des Buches "Obersorbisch – Selbststudium" (Lektion 02, pp12-13).

Nach der Auflistung der Grapheme und der Phoneme im X-SAMPA-Format wurde für jedes Phonem die beste Approximation durch ein deutsches Phonem gesucht. Mithilfe von Ausspracheregeln konnten aus den Graphemen Phonemfolgen erzeugt und das deutsche Akustikmodell angepasst werden.

2.1 Definition des Graphem-Phonemformats

Die Grapheme und Phoneme sind in Tabelle 1 dargestellt. Diese werden nachfolgend einheitlich in Großbuchstaben konvertiert. Die Phoneme werden im X-SAMPA-Format präsentiert und auf den deutschen Phonembestand des UASR-Frameworks (Unified Approach to Signal Synthesis and Recognition) abgebildet. Letzteres besteht aus 43 Einheiten. Eine vollständige Beschreibung befindet ist unter folgender Adresse zu finden:

https://rawgit.com/matthias-wolff/UASR/master/manual/reference/UasrPhonemeSets.html

Jedes Phonem wird außerdem durch ein Aussprachebeispiel in deutscher Sprache ergänzt.

GRPH	X-SAMPA	UASR	Aussprachebeispiel deutsch
A	a	a	A
B	b	b	B
C	ts	t s	Z
Č	tS	t S	TSCH
Ć	t_s	t S	TSCH
D	d	d	D
E	E	E	E
Ě	il	I	E
F	f	f	F
G	g	g	G
H	h	h	[H] (nur vor Vokalen)
I	i	i:	I
J	j	j	J
K	k	k	K
Ł	w	U v	U
L	l	l	L
M	m	m	M
N	n	n	N
Ń	J	j n	JN
O	o	O	O
Ó	uU	U	wie kurzes U
P	p	p	P
R	r	r	R
Ř	S	S	SCH (nur nach p, k und t), nach t auch als Z
S	s	s	ß
Š	S	S	SCH
T	t	t	T
U	u	u:	U
W	v	U v	v (am Anfang und Ende stumm)
Y	1	Y	I
Z	z	z	S
Ž	Z	S	SCH
CH	x	x	CH, am Anfang als KH
Dź	d_Z	d S	DSCH

Fig. 1 Graphem- und Phonembestand.

2.2 Definition der Ausspracheregeln

Die Darstellung der Grapheme im X-SAMPA-Format ist nahezu eineindeutig, während bei der Abbildung in den UASR-Bestand mehrere Grapheme auf dasselbe Phonem abgebildet werden.
Aussprachevarianten bestimmter Graphemsequenzen werden durch Ausnahmen repräsentiert, die in Tabelle 2 zusammengefasst sind.

Regel	Aussprache	Regel	Aussprache	Wortbeispiel	Aussprache
#C_W_$	*	T_Ř_I	t s	ABO	a b O
#C_Ł_$	*	T_Ř_Ě	t s	AFRICE	a f r i: ts E
$_CH_	k	U_Š_Ł	j S	AFRIKA	a f r i: k a
$_CH_C	x	_B_$	p	AFROAMERISKEJE	a f r O a m E r i: s k e: j E
$_H_#V	h	_D_$	t	AGRARNU	a g r a r n u:
$_W_#C	*	_DŹ_$	t S	AKADEMIJE	a k a d E m i: j E
$_W_J	U v	_E_DŹ	e:	AKCEPTOWAĆ	a k ts E p t O U v a t S
$_Ł_#C	*	_E_J	e:	AKCIJE	a k ts i: j E
A_Š_Ł	j S	_E_Ć	e:	AKTERAMI	a k t E r a m i:
E_CH_	C	_E_Č	e:	AKTIWITACH	a k t i: U v i: t a x
I_CH_	C	_E_Ń	e:	AKTIWITAMI	a k t i: U v i: t a m i:
I_J_$	*	_E_Ž	e:	AKTIWITY	a k t i: U v i: t Y
K_Ń_$	*	_H_#C	*	AKTUALNJE	a k t u: a l n j E
S_Ń_$	*	_H_$	*	ALE	a l E
_Ž_$	S	_N_K	n g	ALOJSA	a l O j s a
Ě_CH_	C			AMERIKU	a m E r i: k u:

Fig. 2 Ausspracheregeln (links) und Beispiele.

Der Graphemkontext bestimmt die Phoneme oder deren Auslassung. Die folgenden Symbole wurden dabei für die Definition des Kontextes (<Links>_<Graphem>_<Rechts>) und dessen Aussprache benutzt:

- #C - Konsonant
- #V - Vokal
- $ - Wortgrenze
- * - Auslassung

2.3 Darstellung obersorbischer Phoneme mittels deutscher

Mithilfe der Abbildung and Ausspracheregeln wurde ein Skript zur automatischen Konversion eines Textkorpus in normalisierten Text samt Lexikon entwickelt. Dieses Python-Skript (`corpusProcess.py`) erwartet als Argumente

- den Textkorpus (Audio-IDs) und
- die Phonemabbildung

und generiert Dateien mit folgendem Inhalt (eine Datei pro Stichpunkt):

- normalisierter Korpus (in Großbuchstaben)
- Transliterationsdateien (Audio-IDs)

- Vokabular und Lexikon
- Optional: FSG-Lexikon (finite-state grammar) mit semantischer Abbildung

3 AP 2: Spezifizierung der Sprachanwedung

Eines der Ziele dieser Studie ist es, die Spracherkennung anhand einer praktischen Anwendung auf einem eingeschränkten Sprachdomäne zu demonstrieren. Für diesen Zweck wurde eine Ontologie zur Sprachsteuerung einer Smarten Lampe entwickelt. Realistische Beispiele werden anhand von Satzvorlagen definiert, die sowohl Absichten (intents) mit Parametern als auch „blumige" Phrasen enthalten.

- Definition der Sprachanwendung - Smarte Lampe (SL):
  - Intent 1: Einschalten/Ausschalten
  - Intent 2: Helligkeit setzen: 0-100%, heller, dunkler, dunkel, hell, …
  - Intent 3: Farbe wählen: weiß, rot, gelb, grün, blau, …
- Beispielsätze für die SL-Anwendung:
  - Blumige Phrasen mit Parametern (e.g.):
    "Liebe Lampe, bitte geh <AN>",
    "Bitte setze die Farbe auf <MAGENTA>",
    "Bitte, setze die Helligkeit auf <50%>."

Die Spezifikation wurde in eine BNF-Grammatik (Backus-Naur-Form) umgewandelt, die ihrerseits zur Erstellung einer bestimmten Anzahl zufälliger Sätze im SL-Domäne diente.

Spezifikation

```
\<Lampa\>, daj <#number-hsb\> \<procentow\>.
\<Lampa\>, <#number-hsb\> \<procentow\>.
\<Lampa\>, <#number-hsb\> \<procentow\>, prošu.
…
\<Swěca\>, daj <#number-hsb\> \<procentow\>.
\<Swěca\>, <#number-hsb\> \<procentow\>.
\<Swěca\>, <#number-hsb\> \<procentow\>, prošu.
…
| 030       | třiceći/třicći/třicci
| 035       | pjećatřiceći/-třicći/-třicci
| 040       | štyrceći/štyrcći/štyrcci
| 045       | pjećaštyrceći/-štyrcći/-štyrcci
```

SL-Korpusgenerator

```
{
'<intents>': ['<power_on>','<power_off>','<bright-
ness>','<color>'],
'<power_on>': [
          'Luba <lamp>, zaswěć so.',
          'Prošu, luba <lamp>, zaswěć so.',
          '<lamp> zaswěć so nětko.',
          '<lamp>, Zaswěć so.',
          '<lamp>, dźi on.',
          'Prošu, <lamp>, zaswěć so.',
          'Prošu, <lamp>, dźi on.',
          '<lamp>, trjebam swěcu.',
          '<lamp>, je pře ćma.',
          'Lampa, zaswěćić!',
          'Swěcu, zaswěćić!',
          'Swěcu prošu!',
          'Zaswěć swěcu.',
          'Prošu, zaswěć swěcu.',
          'Trjebam swěcu.',
          'Njewidźu ničo.'
],
…
```

Fig. 3 Spezifikation von Vorlagen und Definition des Korpusgenerators.

Die Minimalmenge der so generierten Sätze wurde schließlich Teil des Satzbestandes für die Sprachaufnahmen.

4 AP 3: Textdatenpräparation

Der validierte Teil der „Common Voice"-Textdaten (CV) ist normalisiert und für die Adaption des deutschen Akustikmodells vorbereitet. Eine Auswahl an CV-Sätzen ist ebenso in die Datensätzen für die Sprachaufnahmen einbezogen, um noch mehr Audiodaten zur Anpassung von Akustikmodellen zu erlangen.
Zusätzlich wurden die CV-Daten für Voruntersuchungen der Performanz des unveränderten deutschen Akustikmodells genutzt (Version 3_20).
Mit den Ergebnissen aus AP1 und AP2 wurden wurde ein Lexikon aus den CV-Daten erstellt. Dieses wurde durch einen Muttersprachler geprüft, die Ausspracheregeln weiter verbessert und ungeeignete Wörter entfernt. Mithilfe des angepassten Lexikons wurden dann ungeeignete Sätze herausgefiltert, die zu Inkonsistenzen in den Phonemadaptionen führen könnten.

Folgende Aufgaben wurden in diesem AP bearbeitet:

- Normalisierung der CV-Daten.
- Extraktion und Validierung des CV-Vokabulars.
- Graphem-Phoneme-Konversion (G2P) auf Grundlage von AP1.
- Erstellung eines Lexikons mithilfe des in Abschnitt 2.3 beschriebenen Skripts.
- Validierung des Lexikons durch einen Muttersprachler.

4.1.1 Common Voice-Korpus

Der CV-Korpus enthält 1600 Audiodateien mit einer Gesamtlänge von 2:42:02.
Sätze mit fremdsprachigen Graphemen und nicht validierten Wörtern wurden ausgelassen, sodass schließlich ein Datensatz mit folgenden Eigenschaften übrigblieb:

- 1352 validierte Sätze
- 5579 Lexikoneinträge

Sprecher	weiblich	männlich
SPK1		172
SPK2		9
SPK3		10
SPK4	44	
SPK5		4
SPK6		5
SPK7		1
SPK8	33	
SPK9		9
SPK10		50
SPK11		98
SPK12		56
SPK13		41
SPK14		4
SPK15		11
SPK16		4
SPK17		808

Fig. 4 Sprecher und Anzahl gelesener Sätze im CV-Datensatz.

Zum Abschluss werden noch einige Beispiele von Sätzen gegeben, die aufgrund von Wörtern und Graphemen fremder Herkunft für die Adaption des akustischen Modells ungeeignet sind.

{'Ź'} NJEDAWNO BU W DELNJEJ ŁUŽICY SERBSKE TOWARSTWO SWÓJBOW GROMAŹE ZAŁOŽENE
{'Ö'} PETRA KÖPPING ŽADA SEJ PŘIPÓZNAĆE ŽIWJENSKEHO WUKONA WUCHODNYCH
{'Ü'} PO MATURJE NAWUKNY WÓN RJEMJESŁO ČASNIKARJA W GLASHÜTTE
{'Ö'} RHÖN SU SRJEDŹNE HORINY Z NJELIČOMNYMI WULKOTNYMI PUĆOWANSKIMI A KOLESOWARSKIMI ŠĆEŽKAMI
{'Ň', 'V', 'Á'} SŁOWAKSKU SKUPINU BOBÁŇOVCI Z MYTOM WUZNAMJENIĆ BĚ SPONTANY ROZSUD
{'V'} TO STEJ HIŽO MJENOWANEJ A REGION BREMERHAVEN
{'Ś'} W DELNJEJ ŁUŽICY ZARJADOWA MAŚICA SERBSKA NARODNE DRASTOWE SWJEDŹENJE
{'X', 'Q'} ZA TO SKIĆI DŹĚĆACE WOČERSTWJENIŠĆO QUERXENLAND WE WODOWYCH HENDRICHECACH DOBRE MÓŽNOSĆE

5 AP4: Aufbau der Sprachaufnahmen

Aus der Kombination von CV- und SL-Daten wurden drei Sätze phonetisch ausgewogener Sätze (PBS) extrahiert: HSB-1, HSB-2, HSB-3. Etwa zwei Drittel der Sätze (Teile der CV-Daten) wurden zur Anpassung des Akustikmodells verwendet, der Rest (Teile der SL-Daten) zur Phonemerkennung und Auswertung der Zustandsgrammatik (FSG).
Die Sätze (HSB-x) wurden in eine Aufnahmesoftware importiert, die Hardware eingerichtet und das Gesamtsystem getestet.

Folgende Aufgaben wurden in diesem AP bearbeitet:

- Entwicklung einer Software (Python-Skript) zur Extraktion der PBS
- Extraktion von PBS aus den validierten CV-Daten zur Adaption des akustischen Modells
- Import der PBS und SL-Beispiele in die Aufnahmesoftware (BAS - SpeechRecorder)
- Einrichtung der Aufnahmegeräte
- Softwareinstallation und Tests

5.1 Auswahl phonetisch ausgewogener Sätze (PBS)

Ein großer Textkorpus ist nötig, um Statistiken über Phonemeinheiten zu erlangen. Dazu wurde folgender Korpus genutzt: „Monolingual Upper Sorbian Data", Shared Task: Unsupervised MT and Very Low Resource Supervised MT, EMNLP 2020 - FIFTH CONFERENCE ON MACHINE TRANSLATION (WMT20), November 19-20, 2020 (http://www.statmt.org/wmt20/unsup_and_very_low_res/).
Der einsprachige obersorbische Datensatz des Sorbischen Instituts enthält eine Mischung aus einem qualitativ hochwertigen Korpus und einiger Daten mittlerer Qualität. Insgesamt sind 251358 Wörter enthalten.
Aus diesen Daten wurden die Häufigkeiten von Phonemen, Diphonen und Triphonen berechnet, um die die phonetische Reichhaltigkeit der o.g. Datensätze HSB-x und des CV-Datensatzes beurteilen zu können.
Mithilfe eines Bewertungsalgorithmus (*Berry & Fadiga "Data-driven Design of a Sentence List for an Articulatory Speech Corpus*), angewandt auf die Diphone, wurden dann Sätze für die Aufnahmen ausgewählt. Das PBS-Auswahlverhalten wurde evaluiert, indem die Anzahl der phonemischen Einheiten in der Satzauswahl den phonemischen Einheiten einer zufälligen Satzauswahl gegenübergestllt wurde.

Ein sehr wichtiges Kriterium bei der Erstellung der Aufnahmedatensätze ist die Länge der Sätze. Ziel waren mindestens 20 Minuten Sprechzeit pro Sprecher.

```
Set Nr.: 1
sentences: 400
sell word count: 2835
estimated duration by phoneme units (min): 20.097777777777775 - 30.146666666666665
estimated duration by number of words (min): 17.71875 - 28.35

type: phones
total tokens: 29
random ratio: 1.0
rand diff: set()
selected ratio: 1.0
sell diff: set()
```

Fig. 5 Ausgabe der PBS-Extraktion.

Zur Abschätzung der Phonem- und Wortrate des Gesprochenen wurden zwei verschiedene Ansätze verfolgt. Im ersten Ansatz wird davon ausgegangen, dass die mittlere Anzahl von Phonemen pro Sekunde zwischen 10 (Lesen eines Gedichts) und 15 (Sportkommentierung) liegt (*Fonagy, I.; K. Magdics (1960). "Speed of utterance in phrases of different length". Language and Speech. 3 (4): 179–192. doi:10.1177/002383096000300401."*). Der zweite Ansatz basiert auf der Anzahl an gesprochenen Wörtern pro Minute, die zwischen 100 (Englisches BBC) und 160 (Spanisches RNE) liegt (*Rodero, Emma. "A comparative analysis of speech rate and perception in radio bulletins." Text & Talk 32.3 (2012): 391-411*). Mithilfe dieser Zahlen konnte eine voraussichtliche Sprachdauer für die Textdaten ermittelt werden (Tabelle 5).
Anhand dieser Daten wurden schließlich die drei disjunkten Datensätze für die Aufnahmen (HSB-x) generiert.

Einmalige Sätze (~250 aus CV):

- HSB-1: 405
- HSB-2: 404
- HSB-3: 401
- Total: 1201

5.2 Sprachaufnahmegeräte

Jede*r Sprecher*in wurde angewiesen, die Sätze genauso zu lesen, wie sie auf der Anzeige geschrieben waren. Die Gründe dafür, einen solchen Ansatz spontaner Sprache vorzuziehen sind:

- Erfassung von Aussprachevarianten,
- Reduzierung des Aufwandes für die Nachbearbeitung,
- Vermeidung manueller Transkriptionen.

Das aus unserer Sicht am besten geeignete Werkzeug für diese Art von Sprachaufnahmen ist der BAS SpeechRecorder (https://www.bas.uni-muenchen.de/Bas/software/speechrecorder), entwickelt vom Bayrischen Archiv für Sprachsignale des Instituts für Phonetik und Sprachverarbeitung der LMU München.
Die Anzeigen wurden importiert und anonyme Sprecherprofile für jeden Datensatz angelegt.

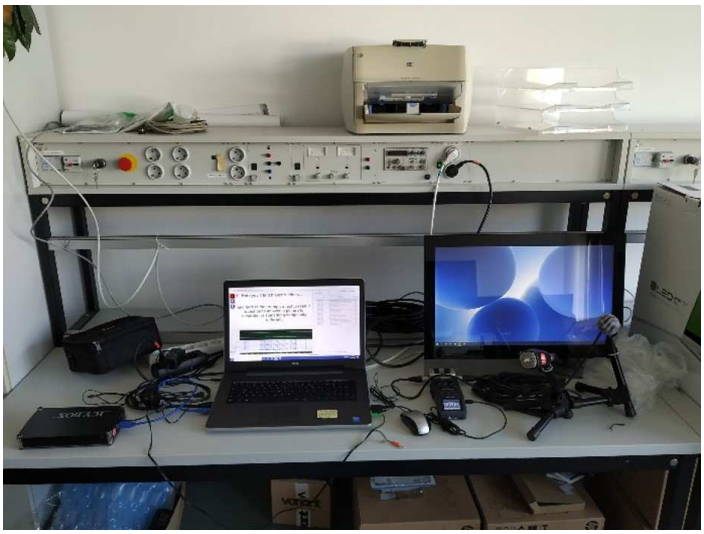


Fig. 6 Geräte für Sprachaufnahmen.
Notebook mit Netzteil,
Zoom H6 Audio-Interface,
EXH-6 Kombikapsel (Gain auf 9, keine Dämpfung),
USB-Kabel für Zoom H6,
Mikrofone (2x Shure SM58, 1x Sennheiser),
Mikrofonstative x 2,
lange XLR-Kabel x 2,
lange HDMI-Kabel,
Monitor,
Kopfhörer,
USB Audio-Interface (Backup),
Externe HDD,
div. Kabel, Netzteile

6 AP5: Organisation der Sprachaufnahmen

Voraussetzung für eine gute Adaption des Akustikmodells sind Sprachaufnahmen von Sprechergruppen, ausgewogen bzgl. Alter und (Sprach-)Geschlecht. Insgesamt wurden 30 Sprecher*innen durch den Auftraggeber angeworben: 10 weibliche, 10 männliche und 10 Kinder. Ein zusätzlicher Sprecher wurde an im Studio der BTU im Rahmen der Testphase aufgenommen.
Hauptgrund für die Aufnahme von Kindern war die Validierung der akustischen Modelle hinsichtlich Phonemerkennung. Die Kinder waren Grundschüler aus höheren Klassen bzw. Mittel- oder Gymnasialschüler niedrigerer Stufen, sodass gute Lesefähigkeiten vorhanden waren.
Für jeden Sprecher war eine Gesamtaufnahmedauer von etwa einer Stunde geplant. Jede Sitzung begann mit einer Vorbereitung aus allgemeiner Einführung, Datenschutz- und Einverständniserklärung. Danach lasen die Beteiligten die Sätze vor. Dem Aufnahmeleiter wohnte ein Muttersprachler bei, der die Aufnahmen begleitete und auf die korrekte Aussprache achtete. Sobald sich ein*e Sprecher*in verlas, wurde der Satz wiederholt. Ausgehend von den Testaufnahmen war bereits offensichtlich, dass es schwer würde, alle Sätze eine der Datensätze (HSB-x) zu lesen (ca. 400). Daher wurden die Sätze für jede*n Sprecher*in zufällig ausgewählt, und sie konnten so viel lesen, wie sie wollten und in der Zeit schafften.
Um eine geeignete Anpassung des akustischen Modells zu erlangen, sollten mindestens drei Stunden Sprachaufnahmen in annähernder Studioqualität gesammelt werden und es sollten nach Möglichkeit alle Sätze mindestens einmal gelesen werden:

- Aufnahmezeitraum: 21.-25.09.2020
- Teilnehmer: 30 (6 pro Tag)
- Sitzungsdauer: ~1 h pro Sprecher
- Aufnahmedauer: 10:09 (hh:mm), 11:30 (inkl. Vortest)

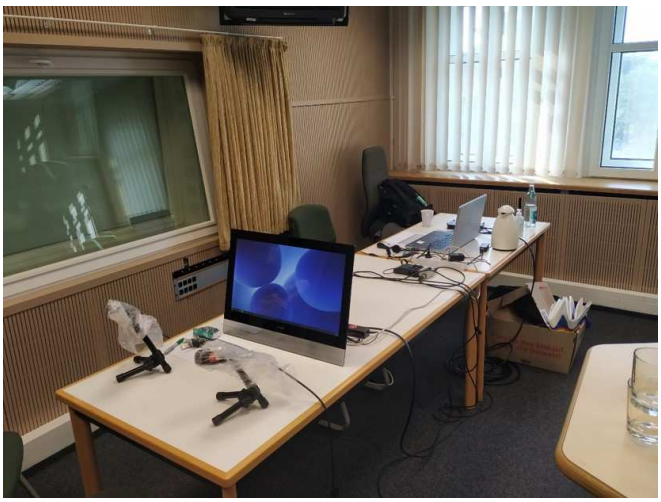


Fig. 7 Aufbau für Sprachaufnahmen.

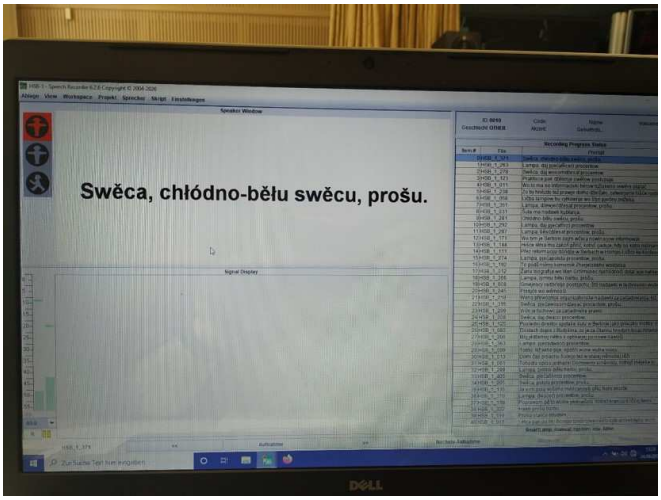


Fig. 8 BAS SpeechRecorder Interface (Bedieneransicht)

Abbildung 9 zeigt die Anzahl der Sätze über alle Datensätze und Sprecherkategorien (M - männlich, F - weiblich, and C - Kinder).

Datensatz	Geschlecht	Sätze	%	Dauer (s)
HSB-1	C	407	7.30%	2,601.23
HSB-1	F	844	15.14%	5,008.34
HSB-1	M	593	10.64%	3,543.44
HSB-1 Total:		*1844*	*33.08%*	*11,153.01*
HSB-2	C	503	9.02%	2,710.39
HSB-2	F	583	10.46%	3,444.36
HSB-2	M	834	14.96%	4,320.75
HSB-2 Total:		*1920*	*34.44%*	*10,475.50*
HSB-3	C	641	11.50%	5,241.64
HSB-3	F	526	9.43%	4,551.13
HSB-3	M	644	11.55%	5,131.87
HSB-3 Total:		*1811*	*32.48%*	*14,924.64*
TOTAL:		**5575**	**100.00%**	**36,553.15**

Fig. 9 Anzahl und Dauer der Sprachaufnahmen.

Die Anzahl der aufgenommenen Sätze variierte über die Sprecherkategorien hinweg. Viele Sätze aus dem CV-Datensatz waren aufgrund ihres Inhalts und Stils für Minderjährige ungeeignet (Zeitungen, politische und religiöse Themen). Die Kinder benötigten mehr Zeit, um sich den Satz vorher leise durchzulesen und ihn dann laut vorzulesen. IM Gegensatz dazu gab es auch professionelle Sprecher aus Radio und Fernsehen, die viel schneller lesen und aussprechen konnten. Die Anzahl an Sätzen pro Sprecher lag zwischen 100 und 250 mit einem Durchschnitt von 191. Es konnte jedoch eine Abdeckung von 100% erreicht werden, sodass jeder Satz mindestens einmal gelesen wurde. Genauer wurden die Sätze zwischen 1 und 10 Mal gelesen, mit einem Durchschnitt von 5.2.

Domäne	Datensatz	Sätze	%	Dauer (s)
ADAPTATION	HSB-1	1356	21.48%	9,636.26
ADAPTATION	HSB-2	1412	22.37%	9,140.28
ADAPTATION	HSB-3	1285	20.35%	11,675.21
	Total	**4053**	**64.20%**	**30,451.74**
SMARTLAMP	HSB-1	737	11.67%	3,192.01
SMARTLAMP	HSB-2	755	11.96%	3,005.26
SMARTLAMP	HSB-3	768	12.17%	5,030.84
	Total	**2260**	**35.80%**	**11,228.11**
TOTAL:		**6313**	**100.00%**	**41,679.85**

Tab. 10 Anzahl und Dauer der aufgenommenen Sätze über alle Domänen und Datensätze.

Tabelle 10 zeigt die Verteilung der Sätze über den die verschiedenen Teildomänen der HSB-Datensätze. Etwa 2/3 der Aufnahmen wurden zur Adaption und1/3 für Tests der SL-Anwendung verwendet. Die Zahlen zeigen eine beinahe Gleichverteilung über alle Datensätze, was unserem Ziel entspricht.

7 Realisierung der Sprachapplikation

Eine der Anforderungen war, lediglich Open-Source-Software und frei verfügbare akustische Modelle für deutsche Phoneme zu verwenden. Eine detaillierte Beschreibung der Entwicklung einer prototypischen automatischen Spracherkennung für Obersorbisch ist in einem separaten Dokument zu finden (Step-by-Step User Guide, Prototypical ASR in Upper Sorbian).

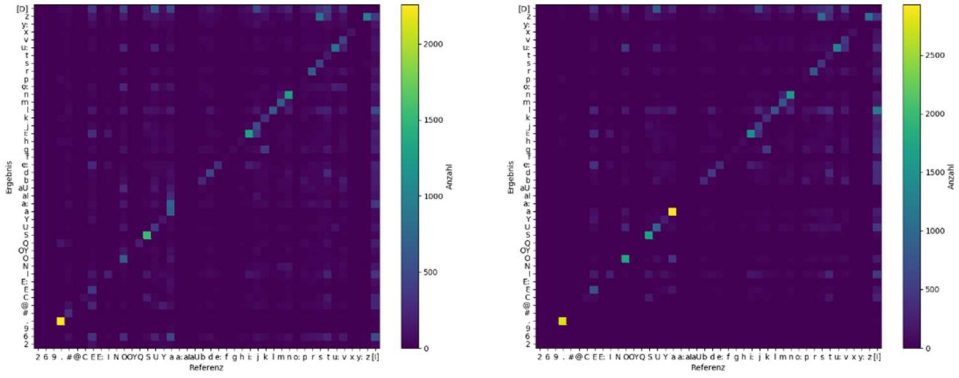
7.1 Nachbearbeitung des Korpus

Die originalen Sprachaufnahmen und Projekteinstellungen wurden gemäß den Datenschutzbestimmungen gesichert und archiviert. Die Aufnahmen wurden nachbearbeitet und im gewünschten Format in eine Datenbank geschrieben: Monokanal Audiodateien, Liste der Dateipfade, Transliteration für jede Aufnahme.

7.2 Phonemerkennung

7.2.1 Optimierung der Phonemabbildungen

Mithilfe der heuristischen Phonemabbildung aus AP1 wurde die Phonemerkennug (PR) auf den CV-Teilen der Datensätze untersucht, um häufige Phonemerkennungsvertauschungen zu finden.



a. Kompletter Phonemsatz b. Reduzierter Phonemsatz

Abb. 11 Vertauschungsmatrizen der Phonemerkennung für die CV-Daten.

Nach der initialen PR-Evaluierung, dargestellt in Abb.11-a, wird deutlich, dass es eine Menge Vertauschungen von Vokalen (<a> mit <a:>, <aI>, <aU> usw.) gibt. Daher wurde die Menge der Phoneme des deutschen Akustikmodells (3_20) auf die in AP1 definierten reduziert (von 43 auf 29). So wird die Anzahl der möglichen Phonemsequenzen reduziert und die Belastbarkeit des Modells gesteigert. In Abb. 11-b ist deutlich zu sehen, dass es weniger Vertauschungen gibt und die Genauigkeit der Phonemerkennung gesteigert wird. Dennoch ist das Ergebnis weit von der idealen Qualität für die CV-Daten entfernt. Die Ergebnisse der Phonemerkennung sind in Tabelle 12 zusammengefasst.

Modell	FA	Corr (%)	Err (%)	LD (%)
3_20	866/1352	30.47	81.51	102.29
3_20_hsb	928/1352	40.34	74.50	104.12

Tab. 12 Ergebnisse der Phonemerkennung.

Aus der Anzahl der Phoneme in der Referenzsequenz (*N*), der Anzahl entnommener Phoneme (*D*), der Anzahl ersetzter Phoneme (*S*) und der Anzahl eingesetzter Phoneme (*I*) wurden die Parameter Korrektheit (*Corr*) und Fehler (*Err*) berechnet. Die Berechnung erfolgte über einen Sequenzabgleich unter Verwendung der Levenstein-Metrik:

$$Corr = (N - D - S) / N$$

$$Err = (D - S - I) / N$$

Die Kennzeichnungsdichte (label density, LD) beschreibt das Verhältnis zwischen der Anzahl an erkannten Phonemen und der Anzahl der Phoneme in der Referenz. Das Forced Alignment (FA) gibt die Anzahl an erfolgreich ausgerichteten Äußerungen an. Es ist offensichtlich, dass eine Reduzierung der Phonemmenge die Genauigkeit erhöht und die Anzahl der Fehler reduziert. Allerdings sind die Erkennungsergebnisse des akustischen Modells noch nicht hinreichend. Um diese zu verbessern, muss das Modell noch an die

Sprache und die akustische Umgebung angepasst werden (z.B. Mikrofonierung, Audioschnittstelle, Raumakustik)

7.3 Akustische Modellierung

7.3.1 Adaptation und Evaluierung

Die Adaption des Modells wurde auf demselben Datensatz durchgeführt, sodass Phonemabbildung und Kennzeichnung (labeling) beibehalten werden konnten. Auf die HSB-Datensätze wurde der Maximum-A-Posteriori-Algorithmus (MAP) angewendet, um die Mittelwerte und Kovarianzen der Gaußschen Verteilungen anzupassen. Die Adaption und Auswertung des akustischen Modells wurde mit der s.g. Leave One Group Out-Kreuzvalidierung (LOGO) vorgenommen. Das bedeutet, dass jeweils zwei der HSB-Datensätze zur Adaption benutzt wurden und das adaptierte Modell auf den verbliebenen Daten getestet wurde.

Beispiel:

- Anpassung an HSB-1 und HSB-2.
- Test auf HSB-3.

Entsprechend werden die Adaptionsversuche mit 12_3, 23_1 und 13_2 bezeichnet.
Die Resultate sind in einem Datensatz zusammengefasst, in dem für jeden erkannten Satz der Sprecher des Satzes selbst nicht Teil der Anpassung war. Schließlich werden alle Datenteilmengen der drei HSB-Datensätze benutzt, um das vollständige und das reduzierte akustische Modell anzupassen, das wiederum zur Evaluierung der FSG benutzt wird:

- 3_20_adp – angepasstes Modell
- 3_20_hsb_adp - angepasstes reduziertes Modell

3_20	**Corr**	**Err**	**LD**	**Dauer**
C	43.8298	68.6058	103.5257	25.09%
F	44.8489	65.8319	101.9966	28.73%
M	41.4213	66.6224	96.6623	46.18%
Total	***42.9510***	***66.8634***	***99.7715***	***100.00%***
3_20_hsb				
C	51.8163	62.0405	104.1836	24.78%
F	51.6274	60.2459	102.2890	29.04%
M	46.8551	62.4525	96.8652	46.17%
Total	***49.3586***	***61.7301***	***100.0989***	***100.00%***
3_20_adp				
C	66.2270	42.1800	98.91	24.41%
F	68.4585	38.5091	98.08	27.96%
M	65.6628	39.7735	94.35	47.64%
Total	***66.5563***	***39.9759***	***96.41***	***100.00%***
3_20_hsb_adp				
C	68.9453	39.5771	98.74	24.33%
F	69.9153	37.3108	98.08	28.08%
M	68.3447	37.4411	94.17	47.59%
Total	***68.9123***	***37.8901***	***96.28***	***100.00%***

Tab. 13 Phonemerkennung nach Anpassung.

7.3.2 Evaluierung des angepassten Modells

Aus Tabelle 13 ist eine absolute Verbesserung der Fehlerraten (Err) nach der Anpassung ersichtlich: Von 66.86 % auf 39.98 % für das vollständige akustische Modell und von 61.73 % auf 37.89 % für das Modell mit reduziertem Phonembestand. In Abb.14 sind die Verwechslungsmatrizen für den Testdatensatz HSB-1 dargestellt (Anpassung an HSB-2 und HSB-3), vor und nach der Anpassung.

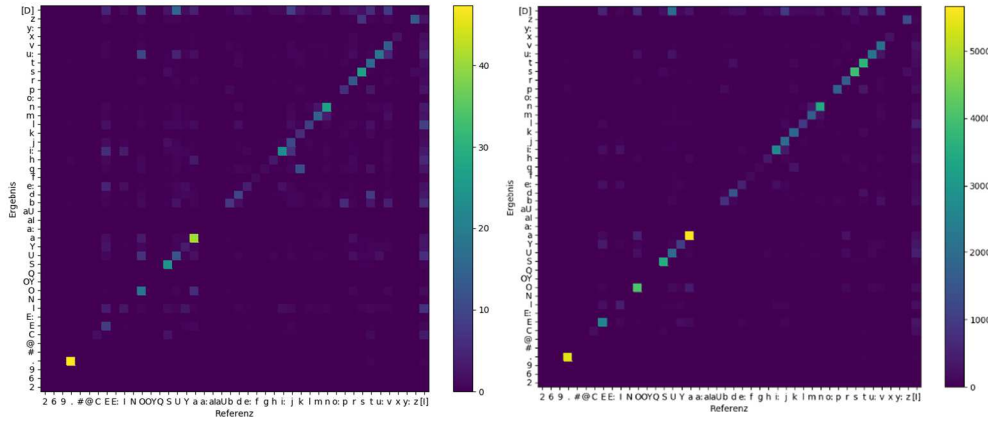


Abb. 14 Phonemverwechslung nach Anpassung (HSB-1)

a. Grundmodell (3_20_hsb) b. Angepasstes Modell (HSB-[23])

Es ist deutlich zu sehen, dass das adaptierte Modell viel weniger Phonemverwechslungen gegenüber dem Grundmodell aufweist (Abb. 14-b). Einige der Verwechslungen bleiben jedoch weiter bestehen. Diese gehören alle zur Gruppe der Verschlusslaute: labiodentale (/b/, /p/), alveolare (/d/) und velare Plosive (/g/).

8 AP 7: Sprachanwendung - Smarte Lampe

Das Sprachmodell für die SL-Anwendung wurde in Form von FSG-Regeln geschrieben. Um Fehler und problematische Regeln zu identifizieren, wurde die Modelle an überangepassten akustischen Modellen getestet und optimiert, um die besten Erkennungsergebnisse zu erzielen. Die Optimierung erfolgte vor allem hinsichtlich folgender Merkmale:

- Aussprachevarianten und Sprechgeschwindigkeit.
- Semantische Kennzeichnung, d.h. Nummer werden demselben Symbol zugeordnet.

```
GRM: <BRIGHTNESS> <LAMP> DAJ:DAJ <PERCENT> PROCENTOW:PROCENTOW
GRM: <BRIGHTNESS> <LAMP> <PERCENT> PROCENTOW:PROCENTOW PROŠU:PROŠU
GRM: <COLOR> <LAMP> <COLHSBACC> <LIGHTACC> PROŠU:PROŠU
GRM: <COLOR> <LAMP> <COLHSBNOM> PROŠU:PROŠU
GRM: <COLOR> <COLHSBACC> <LIGHTACC> PROŠU:PROŠU
GRM: <COLOR> <LAMP> ČIN:ČIN <COLHSBACC> <LIGHTACC>
GRM: <COLOR> <LAMP> POKAZAJ:POKAZAJ <COLHSBACC> <LIGHTACC>
GRM: <COLOR> <LAMP> ČIMOL:ČIMOL <COLHSBACC> BARBU:BARBU
GRM: <COLOR> <COLHSBACC> <LIGHTACC>
GRM: <LAMP>        LAMPA:LAMPA
GRM: <LAMPACC>     LAMPU:LAMPU
```

Abb. 15 Beispiele der FSG-Regeln.

Nach der Optimierung des FSG-Regeln, wurden diese hinsichtlich Wortfehlerraten (word error rate, WER) ausgewertet (Tab. 16). Es stellte sich heraus, dass das reduzierte und LOGO-adaptierte Modell (FSG_HSB_LOGO) die besten Ergebnisse liefert.
Um jedoch einen Sprecherunabhängigen und gegenüber der akustischen Umgebung robusten Demonstrator zu bekommen, wurde das akustische Modell ausschließlich an den CV-Datensatz angepasst und an den HSB-Daten evaluiert (FSG_CV_HSB_LOGO). Eine Verschlechterung der Ergebnisse ist zu sehen, aber gering und es ist zu erwarten, dass das Modell zuverlässig entsprechend der Gesamtqualität des Audiosignals funktioniert.

Tab. 16 WER für die SL-Anwendung.

Modell	Corr	Acc	LD
FSG_DEF_LOGO	93.9	92.8	99.5
FSG_CV_DEF_LOGO	91.0	89.9	98.5
FSG_HSB_LOGO	94.4	93.3	100.0
FSG_CV_HSB_LOGO	92.4	91.3	99.2

9 Zusammenfassung und Ausblick

9.1 Übersicht

Diese Arbeit zeigt, dass es eine sprachübergreifende Anpassung akustischer Modelle mithilfe von Transferlernen im Falle des Obersorbischen möglich ist. Ein bestehendes Modell (deutsch) wurde mithilfe eines großen Sprachkorpus erfolgreich auf den obersorbischen Phonembestand und die akustische Umgebung der Aufnahmen angepasst. Die Auswertung erfolgte auf einer sprecherunabhängigen Phonemerkennung und einer einfachen Kommandozeilensprachanwendung.
Das Ziel der Studie war es, die praktische Anwendbarkeit des Prototyps zu demonstrieren. Es gibt jedoch einige Einschränkungen der Technologie. Um die Belastbarkeit zu verbessern und die Phonemerkennungsraten zu erhöhen, sollten die Modelle an nur einen Sprecher und eine akustische Umgebung angepasst werden (Mikrofonierung, Audioschnittstelle, Raum, Hintergrundgeräusche etc.), d.h. es sollte eine sprecherabhängige automatische Spracherkennung (ASR) eingesetzt werden. Jeder Nutzer des Systems sollte ein Sprecherprofil (akustisches Modell) besitzen, das in einer kurzen Registrierung angelegt wird, indem der*die Sprecher*in eine kleine Anzahl phonetisch ausgewogener Sätze liest. Danach kann das akustische Modell unabhängig von der Sprachdomäne benutzt werden (geringer oder breiter Wortschatz, Ablaufsteuerung, Diktat).

9.2 Kontinuierliche Spracherkennung mit großem Wortschatz

Die folgenden Abschnitte beschreiben in Kürze die Anforderungen an unterschiedliche ASR-Komponenten, um die obersorbische Spracherkennung zu verbessern und sie hin zu sprecherunabhängigen LVCSR-Systemen (large vocabulary continuous speech recognition) mit breitem Wortschatz hin zu erweitern.

9.2.1 Phonembestand

Die Studie ergab, dass es keine passende Eins-zu-eins-Abbildung zwischen den Phonembeständen des Deutschen und des Obersorbischen gibt. Für die folgenden Phoneme ist das klar:

Graphem	IPA	X-SAMPA	UASR
Č	[tʃ]	tS	t S
Ć	[tɕ]	t_s	t S
Ž	[ʒ]	Z	S
Dź	[dʑ]	d_z\	d S

Diese werden als Kombinationen einzelner Phoneme modelliert, wie etwa im Fall von Č und Ć, die beide als Kombination von t und S ausgedrückt werden, die ihrerseits den Graphemen T und Š ensprechen. Auf diese Weise konnte die Anzahl der benötigten Phonemmodelle von 43 auf 29 reduzieret werden, während gleichzeitig die Qualität einiger Phonemmodelle sank (/S/).

9.2.2 Akustische Modellierung

9.2.2.1 Verdecktes Markovmodell (HMM) und Gaußsches Mischmodell (GMM)

Anstatt verschiedene Phonembestände aufeinander abzubilden (eins-zu-eins, eins-zu-viele, viele-zu-viele), erhält man bessere akustische Modelle, indem man die Phonemmodelle selbst modifiziert. Verbesserungen sind zu erwarten, indem neue Phonemmodelle mit eindeutiger Kennzeichnung im X-SAMPA-Format erstellt werden:

- Direkte Verschmelzung zweier oder mehrerer Modelle zu einem.
- Duplizierung eines Phonemmodells.
- Einfügen neuer, leerer Modelle.

Die Anpassung (oder das Umschulen) des neu erstellten Modells erfolgt anhand eines aufgenommenen Korpus, unter der Annahme hinreichender Realisierungen der seltenen Phoneme für eine zuverlässige akustische Modellierung.
Akustische Modelle (HMM und GMM) für Triphone unterscheiden sich von Sprache zu Sprache. Hier sind weitere Anpassungen nötig, z.B. mit weiteren Transduktoren (Finite-State Transducer, FST) oder Abbildungen basierende auf neuronalen Netzen.

9.2.2.2 Tiefe neuronale Netze (DNN)

Akustische Modelle auf Basis tiefer neuronaler Netze (Deep Neural Networks, DNN) sind im Allgemeinen schwerer zu trainieren und anzupassen als die traditionellen HMM/GMM-Modelle, da viele Daten benötigt werden. Andererseits verbessern Aufnahmen weiterer Sprecher und verschiedener akustischer Umgebungen die Belastbarkeit der Modelle gegenüber Variationen (verschiedene Sprecher und Hintergrundgeräusche). Hier sind Ergebnisse auf dem aktuellen Stand der Technik zu erwarten, vorausgesetzt, die Big-Data-Anforderungen sind erfüllt. Das betrifft nicht nur Sprach- sondern auch Textdaten.
Die sprachübergreifende DNN-Sprachdomänenadaption wird für gewöhnlich unter Zuhilfenahme eines existierenden multilingualen DNN-Modells durchgeführt, das auf mehreren Sprachen derselben Sprachfamilie (z.B. westslavische Sprachen) trainiert wurde.
Ein anderer Ansatz besteht darin, verschiedene akustische Modelle verwandter oder unverwandter Sprachen parallel zu verwenden.
Eine Übersicht zu den Herausforderungen in Sprachübergreifender Spracherkennung ist in [*Laurent Besacier, Etienne Barnard, Alexey Karpov, Tanja Schultz, Automatic speech*

recognition for under-resourced languages: A survey, Speech Communication, Volume 56, 2014, Pages 85-100, ISSN 0167-6393] zu finden.

9.2.3 Lexikon-Modellierung

In dieser Studie wurden wissensbasierte, manuell erstellte Regeln auf Basis von Graphemen für die Graphem-Phonem-Sequenzabbildung genutzt. Das so erstellte Lexikon profitiert von Aussprachevarianten, die in der alltäglichen Kommunikation obersorbischer Muttersprachler zu finden sind. (z.B. PROŠU" gesprochen als P'OŠ').

9.2.3.1 Verbundbasierte Methoden (Joint-sequence models)

Ein fortgeschrittenerer, datengetriebener Ansatz ist die Modellierung von Abbildungen ganzer Sequenzen mithilfe statistischer maschineller Übersetzungen, z.B.:

- *Sequitur G2P*, [M. Bisani and H. Ney. "Joint-Sequence Models for Grapheme-to-Phoneme Conversion". Speech Communication, Volume 50, Issue 5, May 2008, Pages 434-451]
- PhonetisaurusG2P, Weighted Finite-State Transducers (NOVAK, J., MINEMATSU, N., & HIROSE, K. (2016). Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering, 22*(6), 907-938. doi:10.1017/S1351324915000315)

Dieser Ansatz erfordert einen Abgleich von Graphemen und Phonemen, wobei die einzelnen Grapheme nicht eins-zu-eins den Phonemen entsprechen müssen.

9.2.3.2 Methoden auf Basis neuronaler Netze

Erst seit Kurzem gibt es Ansätze basierend auf rekurrenten neuronalen Netzen (RNN), die, ähnlich dem vorangegangenen Ansatz, in der Lage sind, ganze Graphemsequenzen in Phonemsequenzen in Abhängigkeit vom Gesamtgraphemkontext zu übersetzen:

- RNN-LSTM (K. Rao, F. Peng, H. Sak and F. Beaufays, "Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, 2015, pp. 4225-4229, doi: 10.1109/ICASSP.2015.7178767) or
- RNN-BLSTM (Mousa, Amr El-Desoky, and Björn W. Schuller. "Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks for Grapheme-to-Phoneme Conversion Utilizing Complex Many-to-Many Alignments." *Interspeech*. 2016.)

9.2.4 Sparachmodellierung

Abhängig von der beabsichtigten Anwendung kann das Sprachmodell anhand einer kontextfreien Grammatik (CFG) oder mithilfe statistischer Sprachmodellierung (SLM) definiert werden. CFG eignen sich für eingeschränkte Anwendungsdomänen mit geringem Wortschatz (einige hundert bis tausend Wörter), in denen sich die Äußerungen nach den erwarteten Wortfolgen richten. SLM hingegen schätzt die Wahrscheinlichkeit einer Wirtfolge basierend auf einer n-Gramm-Statistik (Monogramm, Bigramm usw.). In LVCSR-Systemen werden für gewöhnlich Back-off-Trigramm-Modelle eingesetzt. Abhängig von der Einsatzdomäne kann die Anzahl an Vokabeln hier mehrere hunderttausend Wörter erreichen.
Für das Training eines SLM sind sehr große Mengen an Text erforderlich. Einige Beispiele für Zieldomänen sind Nachrichten, Enzyklopädien, Medizinanwendungen, Rechtstexte oder technische Schriften. SLMs können beliebige Wortfolgen erkennen, sofern die einzelnen Wörter Bestandteil des Vokabulars sind. Um aber optimale Ergebnisse hinsichtlich WER zu bekommen, sollte das Vokabular nach Erscheinung im Quellkorpus

eingeschränkt werden, sodass etwa nur die N häufigsten Wörter (z.B. N=50000) oder nur Wörter mit einer Mindesthäufigkeit (z.B. min. fünfmal erschienen) benutzt werden. Die besten Ergebnisse erreicht man auch hier, indem man sowohl das akustische Modell als auch das Sprachmuster an den jeweiligen Sprecher anpasst. Die Anpassung der Sprachmuster erfolgt dabei durch Aktualisierung des größeren Modells mithilfe der gewichteten Häufigkeiten der N-Gramme, die auf einem kleineren Teil des Textes berechnet wurden. Die neuen, außerhalb des Wortschatzes liegenden Wörter müssen dann ins Lexikon zusammen mit der richtigen Aussprache eingetragen werden.

9.2.5 Semantische Modellierung

In vielen Fällen ist es schwierig, optimale Ergebnisse bei der LVCSR zu erzielen, insbesondere in sprecherunabhängigen Systemen. In vielen Anwendungen abseits von Diktiersystemen, etwa für Medizin- oder Rechtsprotokolle, ist es wichtiger, den Sinn des Gesprochenen zu erkennen. Kleine Fehler bei der Erkennung sind dort meist zulässig.
Das Sprachverständnis (Natural Language Understanding, NLU) liefert Absichten und Objekte aus der erkannten Sprache als Folge von Wörtern. Semantische Markierungen können direkt in das Sprachmodell selbst mithilfe eines Wortklassenmodells eingebunden werden, sie können als Regelsatz in Form einer CFG definiert oder statistisch über einen semantisch gekennzeichneten Textkorpus modelliert werden.

9.3 Zusammenfassung

In Anbetracht der erwähnten Punkte sowie der Ergebnisse dieser Studie wäre ein obersorbisches LVCSR-System unter folgenden Bedingungen praktisch realisierbar:

- Verbesserug der Phonemabbildungen und -modelle
- Erweiterung der G2P-Regeln um Aussprachevarianten
- Anpassung der akustischen Modelle:
  - Sprecherunabhängige ASR: hohe Anzahl an Sprechdaten vieler verschiedener Sprecher und akustischer Umgebungen
  - Sprecherabhängige ASR: Sprecherprofile durch Nutzerregistrierung (Lesen kurzer Texte mit Adpationssätzen)
- Definition der Anwendungsdomäne
- Sammlung domänenspezifischer Texte
- Definition des Vokabulars
- Erstellung eines Lexikons
- Trigramm Sprachmodellierung (SLM)
- Evaluierung der Perplexität des SLM
- WER Perfomance der LVCSR

1 0 Modelltransfer ins Niedersorbische

Da beide Sprachen den Großteil ihres Phonembestands teilen, sind keine größeren Probleme bei der Anwendung des vorhandenen Modells auf das Niedersorbischer zu erwarten. Die folgenden Bedingungen müssen jedoch erfüllt sein:

1. Kompensation der Unterschiede in den Phonembeständen durch entsprechende Phonemabbildungen
2. Anpassung der G2P-Regeln an die Grapheme und Ausspracheregeln
3. Anpassung des Lexikons und der Satzregeln für die FSG der SL-Anwendung
4. Kleiner niedersorbischer Textkorpus für Tests and Evaluation