Fraunhofer
IKTS

FRAUNHOFER INSTITUTE FOR CERAMIC TECHNOLOGIES AND SYSTEMS IKTS

FEASIBILITY STUDY

on Automatic Speech Recognition in Upper Sorbian Language

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

Fraunhofer Institute for Ceramic Technologies and Systems IKTS
Maria-Reiche-Straße 2 01109 Dresden, Germany

**Brandenburg University of Technology Cottbus–Senftenberg,*
Cottbus, Germany

Customer: Stiftung für das sorbische Volk, Postplatz 2, 02625 Bautzen
Offer number: 068-A20-0088-346
Order from: 28.07.2020
Our reference: 068-200025
Report number: 01

Confidentiality: open

Dresden, 2020.11.30

External release

Head of department

Project manager

Table of contents

1 Introduction

This report aims to present the results of the feasibility study for automatic speech recognition in the Upper Sorbian language.

The main objective of the study was to investigate possibilities for transfer learning by using existing language resources (German), in the case of Upper Sorbian (ISO language code: hsb) as an example of an endangered and under-resourced (UR) language.

This was achieved by the development of a voice application demonstrator, a prototypical Automatic Speech Recognition (ASR) system on a limited language domain (Smart Lamp). The report is organized into sections corresponding to the work packages, each starting with a brief introduction of the objectives followed by details of the performed activities. All the involved persons contributed to the successful finalization of the feasibility study.

2 WP1: Grapheme and Phoneme Inventory

The pre-requisite for this work package is a precise definition of the grapheme and the phoneme inventory.

The sources of information which were used: the Wikipedia site for the Upper Sorbian (https://en.wikipedia.org/wiki/Upper_Sorbian_language), the HSB language site of the Sorbian Institute (https://www.obersorbisch.de) and the excerpt of the book “Obersorbisch- Selbststudium” (Lektion 02, pp12-13).

After grapheme inventory and converting phonemes into X-SAMPA format, the Upper Sorbian phonemes were mapped to the nearest German equivalents.
By defining pronunciation rules, it was possible to map the grapheme into corresponding phoneme sequences as a prerequisite German acoustic model adaptation.

2.1 Upper Sorbian to German Phonemes Mappings

The graphemes and the phonemes are shown in Table 1. By internal convention, the graphemes are converted in uppercase format.
The phonemes are presented with X-SAMPA encoding and mapped to German phoneme inventory in the Unified Approach to Signal Synthesis and Recognition (UASR) framework.
The UASR phoneme inventory consists of 43 units. The full description of the phoneme set is given here:

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

For each phoneme, an example of the German pronunciation of the phoneme is given as well.

Fig. 1 Grapheme and phoneme inventory.

GRPH	X-SAMPA	UASR	German Spoken
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	iI	I	E
F	f	f	F
G	g	g	G
H	h	h	[H] (only in front of vowels)
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	short U
P	p	p	P
R	r	r	R
Ř	S	S	SCH (only after p k and t), after t also as Z
S	s	s	ß
Š	S	S	SCH
T	t	t	T
U	u	u:	U
W	v	U v	v (omitted on begin and end)
Y	1	Y	I
Z	z	z	S
Ž	Z	S	SCH
CH	x	x	CH, on the begin as KH
DŹ	d_Z	d S	DSCH

2.2 Definition of Pronunciation Rules

The grapheme to X-SAMPA phoneme mappings are simple “one-to-one” rules, while in mapping to UASR inventory there are some “one-to-many” and “many-to-one” rules. Pronunciation variants of grapheme sequences that compose a word are defined by some exceptions, presented in the left columns of Table 2.

rule		rule		Word (example)	Pronunciation
#C_W_$	*	T_Ř_I	t s	ABO	a b O
#C_Ł_$	*	T_Ř_Ě	t s	AFRICE	a f r i: t s 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 t s E p t O U v a t S
$_Ł_#C	*	_E_J	e:	AKCIJE	a k t s i: j E
A_Š_Ł	j S	_E_Ć	e:	AKTERAMI	a k t E r a m i:
E_CH_	C	_E_Č	e:	AKTIWITACH	a k t i: U v i: t a x
I_CH_	C	_E_Ń	e:	AKTIWITAMI	a k t i: U v i: t a m i:
I_J_$	*	_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 Pronunciation rules definition (left) and some examples from the lexicon

The grapheme context defines the pronunciation of phoneme(s) or their omission. The following symbols were used for the definition of the context (L_GRPH_R) and the conditions of a mapping:

- #C - consonant
- #V - vowel
- $ - word boundary
- * - omission

2.3 Grapheme to Phoneme Conversion

Using the mapping and pronunciation rules, a script for automatic conversion of textual corpus to normalized text and the corresponding lexicon was developed.
Python script (`corpusProcess.py`) takes as input arguments:

- text corpus (audio ids)
- phoneme mapping

and generates the output files:

- normalized corpus, all uppercase
- transliteration files (audio ids)
- vocabulary and lexicon
- optional FSG lexicon with semantic mapping

3 WP2: Speech Application Specification

One of the main objectives of this study is a practical demonstration of speech application on a limited domain (voice control). For that purpose, an ontology of a domain for voice control of a smart lamp is described.
Realistic utterance examples are defined by templates that include the intents along with their arguments, as well as flowery phrases.

- Definition of the speech application - Smart Lamp (SL):
  - Intent 1: Power on/off
  - Intent 2: Set brightness: 0-100%, lighter, dimmer, dark, bright…
  - Intent 3: Set color: warm white, soft white, white, daylight white, cool white, red, crimson, salmon, orange, gold, yellow, green, cyan, sky blue, blue, purple, magenta, pink, lavender.
- Some example sentences for the SL application:
  - Flowery phrases, and intents with parameters (e.g.):
    "Dear lamp, please turn <ON>",
    "Please, set the color to <MAGENTA>",
    "Please, set the brightness to <50%>."

The specification is transformed into a BNF (Backus-Naur Form) grammar which was used to randomly generate any number of sentences in the SL domain.

Specification

```
\<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 Corpus Generator

```
{
'<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 Specification of templates and definition for the corpus generator.

The minimal set of the generated sentences were included in the dataset to be recorded.

4 WP3: Text Data Preparation

The validated part of the “Common Voice” (CV) textual data (version: hsb_2h_2020-06-22) are normalized and pre-processed which could be used for adaptation of the German acoustic model.
A selection of CV sentences is also included in the prompt sets used in speech recordings to provide more audio data on the same utterances for adaptation. Additionally, the CV audio data were used for preliminary performance tests on the default German acoustic model (version: 3_20).
The outcomes from the packages WP1 and WP2 were employed to create the lexicon on the CV data.
The lexicon was checked by a native speaker and the pronunciation rules were further improved and non-suitable words removed.
This lexicon was used to filter out inappropriate sentences which could lead to inconsistencies in phoneme model adaptation.

The tasks fulfilled in this work package are:

- Normalization of Common Voice (CV) textual data.
- Extraction and validation of the vocabulary using CV.
- Grapheme-phoneme conversion (G2P) based on the outcomes from WP1.
- Generation of a lexicon using the tool from Section 2.3.
- Validation of the lexicon by a native speaker.

4.1.1 Common Voice (CV) Corpus

The provided corpus contains 1600 audio files, with a total duration of 2:42:02 (hh:mm:ss). The sentences with foreign graphemes and non-validated words were omitted, providing:

- 1352 validated sentences,
- 5579 lexicon entries.

Speaker	Female	Male
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 Speakers and sentence counts in the CV dataset

Examples of sentences not suitable for acoustic model adaptation, due to words and graphemes of foreign origin.

{'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

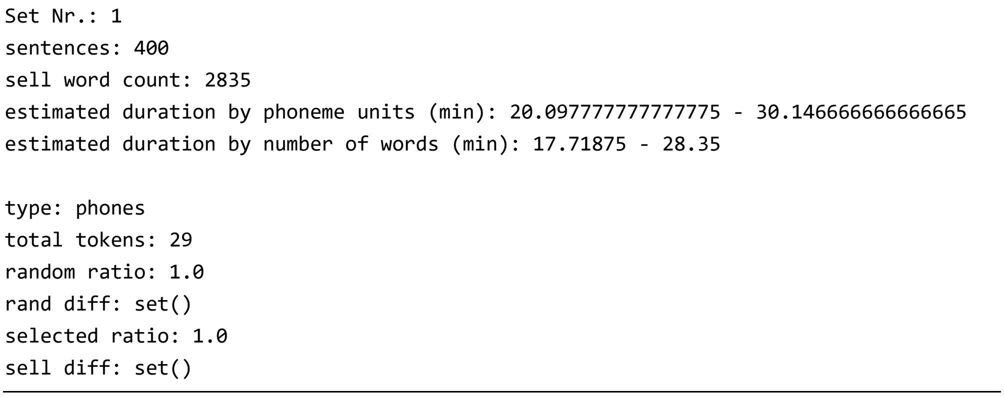
5 WP4: Setup for Speech Recording Sessions

Phonetically Balanced Sentences (PBS) were extracted from the CV dataset and combined with the SL sentences to form the recording prompt sets: HSB-1, HSB-2, and HSB-3.
Around 2/3 of a prompt set is used for acoustic model adaptation (parts of CV) and the rest (parts from the generated SL texts) for phoneme recognition and FSG (Finite-State-Grammar) grammar evaluation. The prompt sets were imported into the recording software, the hardware was set up, and the entire system functionality tested.
The tasks performed in this WP were:

- Development of software tool (python script) for extraction of PBS.
- Extracting PBS from the CV recorded (validated) data for acoustic model adaptation.
- Import PBS and the SL examples into the recording software (BAS - SpeechRecorder).
- Recording equipment setup.
- Software installation and testing.

5.1 Selection of Phoneme Balanced Sentences

A larger textual corpus is necessary to estimate the phonemic units' statistics.
For that purpose we used the corpus: "*Monolingual Upper Sorbian Data*" *from the "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/).
The Upper Sorbian monolingual data (from the Sorbian Institute) contains a high-quality corpus and some medium quality data which are mixed. The vocabulary contains 251358 words.
Using this data, the frequencies of the phones, diphones, and triphones were calculated and used for scoring of the phonetical richness of the sentences in the HSB-1, -2, -3, and the CV dataset.
The sentences were selected according to the scoring algorithm presented in "*Berry & Fadiga* "*Data-driven Design of a Sentence List for an Articulatory Speech Corpus*" applied on the diphones.
The PBS selection performance was compared in terms of the ratio of phonemic units seen in the corpus used for initial statistics (PBS selected against randomly selected).
A very important criterion in designing recording prompts is the estimated duration of spoken sentences. The target was to achieve at least 20 minutes of recorded speech per participant.



```
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 Output from the PBS extraction tool

To estimate the phoneme and word rate in spoken utterance we used two different approaches: "*Fonagy, I.; K. Magdics (1960). "Speed of utterance in phrases of different length". Language and Speech. 3 (4): 179–192. doi:10.1177/002383096000300401".*
The phoneme rate (phonemes per second) was expected to be in the range of 10 (reading poetry) to 15 (commenting sport).
Another approach was based on the word rate (words per minute) presented in *Rodero, Emma. "A comparative analysis of speech rate and perception in radio bulletins." Text & Talk 32.3 (2012): 391-411.)* estimating the values in the range from 100 (English BBC) to 160 (Spanish RNE). Using these figures, we were able to estimate the expected speech duration range by the word and phoneme counts in the selected textual data (see Table 5). The result was that from both HSB and CV data we generated non-overlapping prompt sets, assigned to be recorded by different speakers, HSB-1, HSB-2, and HSB-3.

Unique prompts (~250 from CV):

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

5.2 Speech Recording Equipment

Each speaker in the recording session was instructed to read the prompts exactly as they are presented on the display. The reasons to prefer this approach instead of eliciting spontaneous speech are:

- targeting specific pronunciation variants,
- significantly reducing the effort for postprocessing, and
- completely avoid manual transcription of the recordings.

The tool which is most suitable for this task is the BAS SpeechRecorder (https://www.bas.uni-muenchen.de/Bas/software/speechrecorder) developed at the Bavarian Archive for Speech Signals (BAS) of the Institute of Phonetics and Speech Processing at Ludwig-Maximlians-Universität München.
The prompt sets were imported in the recording tool and anonymous speaker profiles were defined for each HSB set.

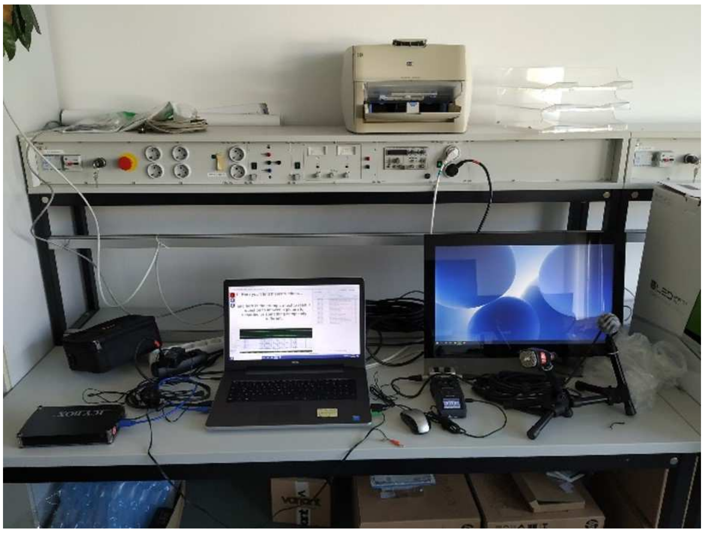


Fig. 6 Equipment used for speech recordings.

Notebook with power adapter
Zoom H6 audio interface and EXH-6 combo capsule (gain set on 9, no attenuation)
USB-cable for Zoom H6
Microphone standers x 2
Long XLR cables x 2
Long HDMI cable
Microphone - Shure SM58 (x 2)
Microphone - Sennheiser x 1
USB audio interface (for backup)
Acer Monitor with power adapter
ICY Box external HDD, cables, power adapter
Extension cord with 3 sockets
Extension cord with 5 sockets
Headphones for audio check

6 WP5: Organization of the Speech Recordings

The prerequisite for good acoustic model adaptation is collecting recordings from gender and age balanced group of speakers.
In total 30 speakers were recruited by the customer: 10 female, 10 male, and 10 children. One additional male speaker was recorded at the BTU studio as a part of the testing phase. The main reason to include children's speech is to see how the adapted acoustic models would perform in terms of phoneme recognition. The children participants are minors attending either higher classes of elementary school or lower classes in high school which implies good reading skills.
For each speaker, a recording session for a duration of one hour was planned. The sessions consist of preparation: general introduction, signing the data protection, and consent waivers. Then, the participants would read the displayed prompts.
Besides the recording operator, a moderator (native speaker) was present for the whole duration of the recording session to ensure proper pronunciation for the speech prompts. Whenever a speaker would mispronounce a word, the prompt recording was repeated. From the experiences in the test recording session, it was obvious that it would be difficult to achieve the maximum number of recordings per prompt set (~400).
Therefore, the order of prompting was randomized, and each speaker was free to read as many as possible.
To achieve proper cross-language acoustic model adaptation we set the target to collect a minimum of 3 hours of speech recordings in near studio quality, also each prompt in the sets must be recorded at least once:

■ Period:	21.-25.09.2020
■ Participants:	30 (6 per day)
■ Session duration:	~ 1 hour per speaker
■ Recorded speech:	10:09 (hh:mm), 11:30 (incl. pre-test)

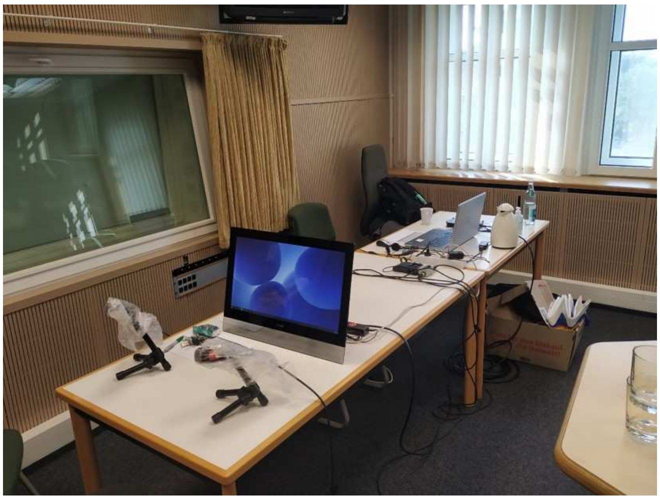
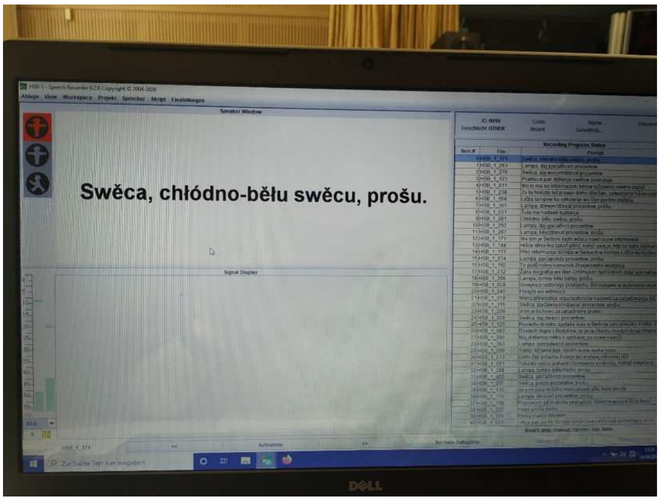


Fig. 7 The recording setup.



Swěca, chłódno-bělu swěcu, prošu.

Fig. 8 BAS SpeechRecorder interface (operator view)

Figure 9 presents the distribution of the number of prompts across datasets and speaker category (M-ale, F-emale, and C-hildren).

Dataset	Gender	Prompts	%	Duration (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 Number and the duration of speech recordings

The number of recorded prompts varied across speaker categories. Many prompts were taken over from the CV data, where the content and the style were on a level not appropriate for minors (newspaper, political and religious texts). The children needed more time, first to read the prompt silently and then to say it. In contrast, some of the participants, professional radio and TV moderators, were able to read and pronounce much faster.
The number of recorded prompts, per speaker, was in the range of 100 to 250 with an average of 191. A coverage of 100% of the available prompts (1210) was achieved ranging from 1 to 10 with an average of 5.2 recordings per prompt.

Domain	Dataset	Prompts	%	Duration (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**

Fig. 10 Number and duration of recorded prompts across datasets and domains

Figure 10 presents the distribution of the prompts across the HSB datasets and the adaptation and the Smart Lamp domain subsets. Around 2/3 of the recordings are considered for adaptation and 1/3 for testing of the Smart Lamp application. Almost equal distribution is achieved across datasets, which was the objective.

7 Realization of the Speech Application

One of the requirements in conducting this study was to use open-source software exclusively, including freely available German acoustic models. A detailed description of the development of the prototypical ASR in Upper Sorbian is given in a separate document “Step-by-Step User Guide, Prototypical ASR in Upper Sorbian”.

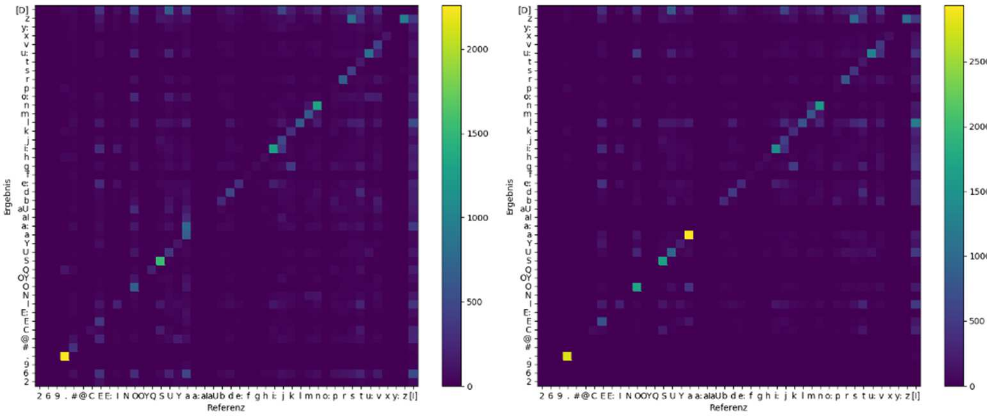
7.1 Corpus Post-Processing

The original recording project setup and the files were secured and archived accordingly to the data protection law and internal regulations.
The recordings were post-processed and converted into a database with the desired signal and metadata format: mono channel audio files, file lists (file paths), corresponding transliterations.

7.2 Phoneme Recognition

7.2.1 Optimization of the Phoneme Mappings

Using knowledge-based phoneme mapping from the WP1, phoneme recognition (PR) performance is evaluated on the CV audio data to investigate frequent phoneme recognition confusions.



a. Full phoneme set b. Reduced phoneme set

Fig. 11 Confusion matrices of phoneme recognition on CV speech data.

After the initial PR evaluation, from the Fig.11-a, it is evident there are many confusions across the vowels (i.e. <a> with <a:>, <al>, <aU> ...). Therefore, the first data-driven based optimization was to reduce the phoneme set of the default German acoustic model (3_20) to only those (3_20_hsb) which were defined in the WP1 (from 43 to 29). This narrows the choice of phoneme sequences and improves the robustness of the model. From Fig-11-b it is obvious that there are fewer confusions with improved phoneme recognition accuracy despite the modest quality of the CV audio data.

The table (Fig. 12) presents the results of phoneme recognition.

Fig. 12 Phoneme recognition performance.

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

The evaluation parameters correctness *(Cor) and* error *(Err)* were calculated using the number of phonemes in the reference sequence *(N)*, the number of removed phonemes *(D)*, the number of substituted phonemes *(S),* and the number of inserted phonemes *(I)*. These numbers are calculated over a sequence alignment using Levenstein distance:

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

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

The label density (LD) indicates the ratio in the phonemes counts of reference and the phoneme recognition. The Forced alignment (FA) presents the number of successfully force-aligned utterances. It is evident that the reduction of the phoneme set improved the correctness and reduced the errors.

However, the performance of the acoustic model is still low, and to improve the performance the model needs to be adapted to the language and the acoustic environment (e.g. microphones, interface, room acoustics).

7.3 Acoustic Modeling

7.3.1 Adaptation and Evaluation

The model adaptation was performed on the same audio datasets keeping the same phoneme mapping and labeling. The maximum a-posteriori (MAP) algorithm was used to adapt the mean and covariance of the Gaussian distributions using the adaptation portion of the HSB datasets.

The adaptation and evaluation of the acoustic models was performed with "Leave One Group Out" cross-validation (LOGO). Namely, two of the HSB sets were used for model adaptation and tested on the third set.
For instance:

- adapt to HSB-1 and HSB-2 and
- test on HSB-3, etc.

Therefore, we denote the adaptation experiments with 12_3, 23_1, and 13_2.
The results are aggregated into one data frame where for each recognized sentence the speaker was not part of the adaptation set.
In the end, we use all the adaption subsets of all 3 datasets to adapt the full and the reduced acoustic model to be used for the FSG grammar evaluation:

- 3_20_adp - adapted model
- 3_20_hsb_adp - adapted reduced model

3_20	Cor (%)	Err (%)	LD (%)	Duration (%)
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%***

Fig. 13 Phoneme recognition after adaptation.

7.3.2 Evaluation of the Adapted Models

From the Table (Fig.13) we can see the absolute improvement in error rates (Err) after adaptation: from **66.86 %** to **39.98 %** for the full acoustic model, and from **61.73 %** to **37.89 %** for the reduced phoneme set model. Figure 14 presents the confusions matrices calculated on HSB-1 as a test set, before and after the model adaptation.

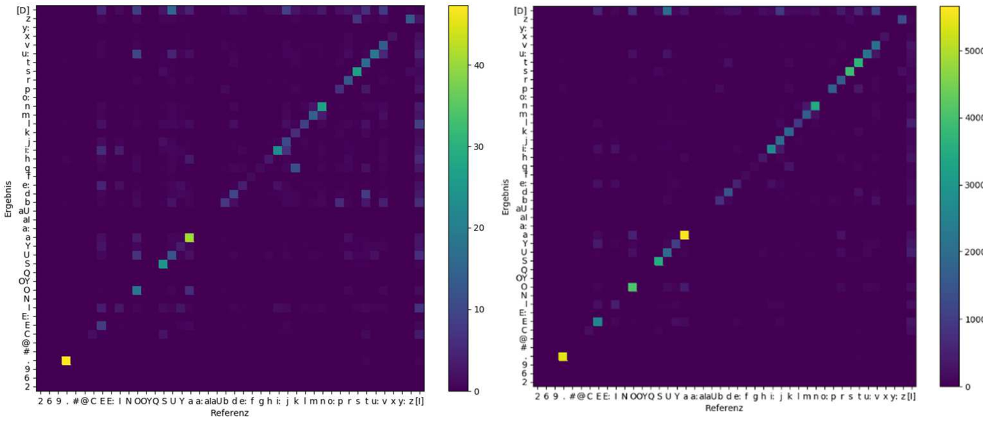


Fig. 14 Phoneme confusions after adaptation (HSB-1)

a. Default model (3_20_hsb)

b. Adapted model (HSB-2 and -3)

There are fewer confusions after adaptation (Fig.14-b). However, it could be noticed that there are still some phoneme confusions. They mostly belong to the group of stop consonants - labiodental plosives: /b/, /p/, alveolar plosive: /d/ and velar plosive /g/.

8 WP7: Smart Lamp Application

The language model for the Smart Lamp application was written in form of Finite-Stats-Grammar (FSG) rules. To identify errors and problematic rules we tested and optimized the language models on overfitted acoustic models to ensure the best recognition performance. The optimization was mostly addressing the:

- Pronunciation variants and speech rate.
- Semantic tagging, i.e. numbers are mapped to the same output symbol.

Fig. 15 Example of FSG grammar rules

```
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
```

After optimizing the grammar, it was evaluated in terms of word recognition accuracy (Acc) performance (Table Fig.16).

It was observed that we achieved the best performance of **93.3%** in the case of HSB reduced and LOGO adapted model (FSG_HSB_LOGO).
However, to deliver a demonstrator that will be speaker-independent and robust to the acoustic environment we adapted the reduced acoustic model exclusively on the CV speech data and test on the HSB datasets (FSG_CV_HSB_LOGO).
The decrease in the performance (**91.3%**) was not drastic and it is expected that the model will perform reliably according to the quality of the audio signal (tested in ad-hoc trials).

Fig. 16 WER on the Smart Lamp Application.

Model	Cor	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 Conclusions and Future Work

9.1 Overview

The feasibility study shows that it is possible to use transfer learning for cross-lingual acoustic model adaptation in the case of Upper Sorbian.
The existing acoustic model (German), trained on a large speech corpus, was successfully adapted on the Upper Sorbian phonetic inventory and acoustic environment of the recordings.
One of the objectives of this study is to demonstrate the practical usability of the voice application. The models were evaluated in speaker-independent phoneme recognition and simple command and control (C&C) voice application (Smart Lamp).
However, there are limitations of the used technology. To improve the robustness and maximize the recognition performance, the acoustic models should be adapted on only one speaker and acoustic environment (microphones, audio-interface, room, background noise, etc.), namely speaker-dependent ASR.
Each user of the system should have a speaker profile (acoustic model) usually created within the enrollment procedure, where the speaker reads a small set of phonetically balanced sentences. Then, the acoustic model can be used regardless of the language domain (small or large vocabulary, C&C, or dictation).

9.2 Large Vocabulary Continuous Speech Recognition

The following sections briefly describe the requirements in different ASR components for improving and expanding the speech technologies in Upper Sorbian, from speaker-dependent and small vocabulary task to speaker-independent LVCSR systems.

9.2.1 Phoneme inventory

From the study, it was observed that there was no suitable one-to-one mapping between the phoneme inventories (Upper Sorbian to German).

This is emphasized for the following phonemes:

Grapheme	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

They are modeled as a sequence of separate phoneme models, like in the case of Č and Ć which are both expressed as a sequence of /t/ and /S/, which also correspond to 2 different graphemes (T and Š).
This reduced the diversity of the phoneme models from 43 to 29 while at the same time degrading the quality of some of the phoneme models (/S/).

9.2.2 Acoustic Modeling

9.2.2.1 HMM/GMM Acoustic Models

Instead of mapping on the level of phoneme inventory (one-to-many, many-to-many, or many-to-one), better acoustic modeling can be done by modifying the phoneme models themselves.
Improvements could be expected by the creation of new phoneme models with unique labels according to the X-SAMPA convention, by:

- direct merging of two or more models into one phoneme model,
- duplicating phoneme model, and,
- inserting new empty models.

Then adaptation (or re-training) of the newly established models is performed on a recorded corpus, under the assumption that there are enough realizations of the rare phonemes for reliable acoustic modeling.
In the case of tri-phone HMM/GMM acoustic models, there will be a mismatch of the triphone sets across the languages. Therefore, appropriate mapping is required (with another FST transducer or neural network-based mapping).

9.2.2.2 DNN Acoustic Models

Deep neural network acoustic models are, in general, more difficult to train and adapt than the traditional HMM/GMM models because of the need for a large amount of speech and language data (big data).
On the other hand, speech data of more speakers and different acoustic environments greatly improve the robustness on speaker and background noise variations.
Here, we can expect State-of-the-Art performance if the requirement of big data is fulfilled. This does not apply only to speech (audio) but also to language data (text).
Cross-lingual DNN language domain adaptation is usually done by leveraging an existing multilingual DNN model, trained on several or more languages of the same or related language family (e.g. West-Slavic group). Another approach is also to use parallel multiple acoustic models of related or un-related languages.
A review of the challenges in cross-lingual speech recognition is given 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*".

9.2.3 Lexicon Modeling

In this study, we used knowledge-based handcrafted rules (grapheme based) for grapheme -to phoneme (G2P) sequences mapping.

The lexicon will benefit from adding more pronunciation variants seen in everyday communication of the native speakers of Upper Sorbian (e.g. “PROŠU" pronounced as “POŠ”).

9.2.3.1 Joint-Sequence Based

A more sophisticated approach (data-driven) is to model the sequence-to-sequence (graphemes - phonemes) by using statistical machine translation principles, like in the tools:

- *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)

This approach requires grapheme-to-phoneme alignments, where graphemes and phonemes do not correspond one-to-one.

9.2.3.2 Neural Network Based

Recently, approaches based on recurrent neural networks (RNN) are introduced. RNNs are capable of translating grapheme sequences into phoneme sequences considering the full context of graphemes.

- 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 Language Modeling

Depending on the intended application the language model can be defined either by Context-Free-Grammar (CFG) rules or by statistical language modeling.
CFG grammars are appropriate for very limited vocabulary (few hundreds to thousand words) where the spoken utterance follow the expected order of words.
In contrast, statistical language modeling (SLM) estimates the probability of word sequences based on N-gram statistics (unigrams, bigrams, trigrams, and more).
Normally, in LVCSR systems a trigram back off models are employed, and depending on the target domain, the vocabulary size could reach several hundred thousand words.
To train SLM, a large amount of in-domain text data is required. Some examples of target domains are news, encyclopedic (Wikipedia), medicine, law, technology, etc.
SLMs can recognize any sequence of words in any order if they are present in the vocabulary. However, to achieve optimal performance in terms of WER, the vocabulary should be restricted by the frequency of word occurrences. Usually, this is done by setting a criterion for a textual corpus, like top-N most frequent words (N=50000), or minimal count (i.e. a word occurred at least 5 times).
State-of-the-art performance with LVCSR is achieved in dictation systems where the acoustic and the language model are adapted to the specific speaker.
Language model adaptation is usually performed by updating the existing larger model with the weighted statistics of the N-grams calculated on a smaller amount of text.

The new Out-Of-Vocabulary (OOV) words should be included in the lexicon with the proper pronunciation.

9.2.5 Semantic Modeling

In most cases, the state-of-the-art LVCSR performance is difficult to achieve, particularly for speaker-independent systems.
However, except in dictation systems (like medical or law protocols transcription), in many applications it is more important to recognize the meaning of the spoken utterance, and misrecognition of some words (e.g. flowery phrases) can be tolerated.
Natural Language Understanding (NLU) component identifies intent(s) and named entities from a sequence of recognized words.
Semantic tags can be directly embedded in the language model itself by word-class modeling, they can be defined as a set of rules in a form of a CFG grammar or statistically modeled over a semantically labeled text corpus.

9.3 Conclusions

LVSCR in Upper Sorbian could by practically feasible under the following conditions:

- Improved phoneme mapping and models.
- Extending G2M rules for pronunciation variants.
- Acoustic model adaptation:
  - speaker-independent ASR: a large amount of speech data with a variety of speakers and acoustic environments.
  - speaker-dependent ASR: speaker profiles by user enrollment (reading a small set of adaptation sentences).
- Definition of the application domain.
- Collection of in-domain textual data.
- Vocabulary definition.
- Lexicon generation.
- 3-gram language modeling (SLM).
- Perplexity evaluation of the SLM.
- WER performance of the LVCSR.

10 Model Transfer to Lower Sorbian

Since both languages share almost the same phonetical inventory, there should be no major issues to employ the adapted acoustic models for recognizing speech in Lower Sorbian.

However, these are requirements to be fulfilled:

1. The differences in the phoneme inventory should be addressed by proper phoneme mapping.
2. Adjust the G2P model to the graphemes and the pronunciation rules.
3. Smart Lamp FSG grammar adapted lexicon and sentence rules.
4. Collection of smaller speech corpus in Lower Sorbian for testing and evaluation.