# CROSS-LINGUAL ACOUSTIC MODELING IN UPPER SORBIAN - PRELIMINARY STUDY

*Ivan Kraljevski*[1], *Marek Rjelka*[1], *Frank Duckhorn*[1], *Constanze Tschöpe*[1], *Matthias Wolff*[2]

[1]*Fraunhofer Institute for Ceramic Technologies and Systems IKTS, Dresden, Germany*
[2]*Chair of Communications Engineering, Brandenburg University of Technology (BTU) Cottbus-Senftenberg, Cottbus, Germany*
*{ivan.kraljevski, marek.rjelka, frank.duckhorn, constanze.tschoepe}@ikts.fraunhofer.de, matthias.wolff@b-tu.de*

**Abstract:** In this paper, we present a preliminary study for acoustic modeling in Upper Sorbian, where a model of German was used in cross-lingual transfer learning. At first, we define the grapheme and phoneme inventories and map the target phonemes from the most similar German source equivalents. Phonetically balanced sentences for the recording prompts were selected from a combination of general and domain-specific textual data. The speech corpora with a total duration of around 11 hours was collected in controlled recording sessions involving an equal number of females, males, and children.
The baseline acoustic model was employed to force-align the speech corpora given the knowledge-based phoneme mappings. How well the mappings were, was evaluated by the phoneme confusions in free-phoneme recognition. The new derived data-driven model with a reduced phoneme set was included in the adaptation and evaluation along with the baseline acoustic model. The model adaptation performance was cross-validated with the "Leave One Group Out" strategy. We observed major improvements in phoneme error rates after adaptation for the knowledge-based and data-driven phoneme mappings. The study confirmed the feasibility of transfer learning for acoustic model adaptation in the case of Upper Sorbian, at the same time demonstrating practical usability with a small vocabulary speech recognition application (Smart Lamp).

## 1 Introduction

In recent times, the adoption of voice assistants has been on the rise worldwide, fueled largely by the ever-improving state-of-the-art (SotA) speech technologies and partially due to people's changed behavior during the pandemic. Many market surveys report that the number of users of voice assistants will grow into double digits in the coming year [1].

Popular voice assistants support languages that have a large percentile of speakers worldwide. One such example is Google Assistant which has around 44 on smartphones and 8 languages on Google Home, with additional dialects, followed by Apple Siri with 21 languages, Amazon Alexa with 8 languages and additional dialects, and Microsoft Cortana with 8 languages [2]. The SotA automatic speech recognition (ASR) systems made a breakthrough in terms of recognition achieving "near-human" performances in restricted conditions, domain, and language. The study [3] shows that current SotA speech recognizers perform poorly in domain-specific use cases under noisy conditions. To achieve high recognition performance in deep neural network (DNN) ASR systems, a huge amount of speech and language data is required to train the models.

The challenges of introducing new languages in SotA ASR systems are multi-fold. It is taken for granted that if enough data for a target language exist or could be collected, then the data amount requirements for reliable speech and language modeling could be fulfilled.

However, in contrast to the 389 languages (nearly 6% of 5000 existing) which covers 94% of the world's population, there are languages with a very low number of native speakers. For them, there is, in general, no economic interest in developing speech applications by the aforementioned big technological giants. Transfer learning could enable rapid development of speech-enabled application on the target language by taking advantage of any available experience, knowledge, and resources (corpus, transcriptions, text, speech, and language models) of the source language.

The simplest approach to transfer learning is to apply model adaptation. Where the adapted model will be used for bootstrapping a new model by increasing the amount of speech with improved transcriptions [4]. However, DNN acoustic models are more difficult to train and adapt due to the required amount of data. Also, because the phonotactic constraints significantly differ across languages from different families [5] it is even more difficult to reuse monolingual DNN models.

Cross-lingual DNN adaptation is usually done by leveraging an existing multilingual DNN model, which is trained on several or more languages of the same or related language family (e.g. West-Slavic group). Another approach is to use multiple acoustic models of related or unrelated languages in parallel. An extensive review of the challenges in cross-lingual speech recognition is given in [4].

In this paper, we present a preliminary study of transfer learning used for rapid development of a speech application in Upper Sorbian. Although the target and the source language (German) belong to different language families, the phoneme inventories share a significant portion of their phonetic units. Therefore cross-lingual acoustic modeling is feasible as a basis for future development. The collected and created resources will be used to bootstrap a SotA speech recognition system with much wider applicability than the developed demonstrator.

## 2 Sorbian Languages

The Sorbian languages (Upper and Lower Sorbian) belong to the West Slavic branch of the Indo-European language family along with Polish, Czech, and Slovak. The Upper Sorbian is spoken in the region of Upper Lusatia and Lower Sorbian is spoken in the region of Lower Lusatia. Although closely related they are only partially mutually intelligible. Both branches kept some special features from Old-Slavic. In general, Upper Sorbian is perceived to be closer to Czech and Slovak languages, whereas Lower Sorbian is considered to be closer to Polish.

Both branches are considered endangered and under-resourced languages. There are no reliable figures about the speakers of Sorbian languages, different sources provide a wide range of estimations which vary between 15 and 30 thousand for Upper Sorbian, and 5 to 10 thousand for Lower Sorbian [6]. Most of the native speakers do not use the language in daily communication, which threatens its existence despite being protected under the European Charter for Regional or Minority Languages.

Recently a substantial effort has been made to preserve the Sorbian languages by various projects conducted by the Foundation for the Sorbian People[1] and by the Sorbian Institute[2]. Other research groups contributed to the collection and documentation of speech and language in Sorbian with particular importance being stressed on digitalization.

The Corpus for Spoken Lower Sorbian (GENIE) [7] combines the audio recordings origi-

[1] https://stiftung.sorben.com

[2] https://www.serbski-institut.de

nating from the archive of Sorbian broadcasts (1956–2006), from the Archive of Sorbian Culture (1951–1971), and new recordings from native speakers (2005–2006). Recently, from 2010-2015 audio corpus in duration of 100 hours was recorded in the Sorbian Institute, Cottbus [8]. The objective was to obtain and document the speech performance of the native speakers in Lower Sorbian. However, to the best of our knowledge, there are no speech corpora in Upper Sorbian that can be feasibly employed for the research and development of speech technologies.

# 3 Grapheme to Phoneme modeling

## 3.1 Knowledge-based Phoneme Mapping

Table 1 presents the differences in phoneme mappings derived from various sources like [9],[10]. For each phoneme, an example of the German pronunciation is given as well. The Upper Sor-

**Table 1** – Grapheme and phoneme inventory

| Grapheme | X-SAMPA | UASR | German Spoken |
|---|---|---|---|
| C | ts | t s | Z |
| Č | tS | t S | TSCH |
| Ć | t_s | t S | TSCH |
| Ě | il | I | E |
| H | h | h | [H] (only in front of vowels) |
| I | i | i: | I |
| Ł | w | U v | U |
| Ń | J | j n | JN |
| O | o | O | O |
| Ó | uU | U | short U |
| Ř | S | S | SCH (only after p k and t), after t also as Z |
| Š | S | S | SCH |
| 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 |

bian phonemes were converted into X-SAMPA format and mapped from the nearest German equivalents. The used German phoneme inventory is defined in the Unified Approach to Signal Synthesis and Recognition (UASR) framework [11] and consists of 43 units [12].

## 3.2 Definition of Pronunciation Rules

The grapheme to X-SAMPA phoneme mappings are simple "one-to-one" rules, while in mapping to UASR phoneme set there are some "one-to-many" and "many-to-one" rules.

Pronunciation variants of grapheme sequences that compose a word are defined by exception rules, presented in the columns of Table 2. The grapheme context (Left_GRPH_Right) defines the pronunciation of phoneme(s) or their omission: "#C" - consonant, "#V" - vowel, "$" - word boundary, and "*" - phoneme omission.

For instance, in the word "zymskich" (/z/ /Y/ /m/ /s/ /k/ /i:/ /C/) the rule "*I_CH_*" was applied, where in "zwučowanjach" (/z/ /U/ /v/ /u:/ /t/ /S/ /O/ /U/ /v/ /a/ /n/ /j/ /a/ /x/) the default grapheme to phoneme rule ("CH" -> "*/x/*") according to Table 1. The pronunciation rules were

**Table 2** – Grapheme to phoneme mapping exceptions

| context | map | context | map | context | map | context | map |
|---|---|---|---|---|---|---|---|
| #C_W_$ | * | T_Ř_I | t s | $_Ł_#C | * | _E_J | e: |
| #C_Ł_$ | * | T_Ř_Ě | t s | A_Š_Ł | j S | _E_Ć | e: |
| $_CH_ | k | U_Š_Ł | j S | E_CH_ | C | _E_Č | e: |
| $_CH_C | x | _B_$ | p | I_CH_ | C | _E_Ń | e: |
| $_H_#V | h | _D_$ | t | I_J_$ | * | _E_Ž | e: |
| $_W_#C | * | _DŹ_$ | t S | K_Ń_$ | * | _H_#C | * |
| $_W_J | U v | _E_DŹ | e: | _Ž_$ | S | _N_K | n g |
| S_Ń_$ | * | _H_$ | * | Ě_CH_ | C | | |

defined according to the examples from [13] (Lesson 02, pages 12-13) and enhanced by a native speaker.

# 4 Speech Data

The speech data originate from two different sources: the validated part of the "Common Voice" (CV) which is a crowd-sourced and open-source dataset, and a speech corpus which was collected in controlled recording sessions (HSB Corpus). Using the defined pronunciation rules we created the lexicon which was used to obtain the phoneme transcriptions, as a prerequisite for the forced-alignment stage in acoustic adaptation.

## 4.1 Common Voice Corpus

The Common Voice [14] dataset (version: hsb_2h_2020-06-22) contains 1600 audio files of 2 female and 15 male speakers with a total duration of 2:42:02. The content originates from various general-domain sources (newspapers, books, proverbs, etc.) The sentences containing graphemes and words of foreign origin were omitted, yielding in total 1352 sentences with 5579 vocabulary entries.

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 modeling.

## 4.2 HSB Corpus

The HSB speech corpus was collected by Fraunhofer IKTS and the Brandenburg University of Technology (BTU) at the premises of the Foundation for the Sorbian People in Bautzen. The recording sessions used three different phonetically balanced prompt sets (HSB-1, 2, and 3). Around 2/3 of the prompts are taken from Common Voice texts and the rest from domain-specific sentences.

The Smart Lamp (SL) data include examples of command and control actions, e.g. turning on/off a smart bulb, setting the brightness and the color. The SL domain is defined by templates with intents and their entities, as well as, flowery phrases. The templates were converted into a BNF (Backus-Naur Form) grammar which was used for random generation of the maximal set for prompting.

### 4.2.1 Prompts Selection

To maximize the benefits of controlled recording sessions, the prompts should be phonetically balanced and rich, resembling the phoneme units' statistics of larger textual data. For that

purpose, we used the corpus "Monolingual Upper Sorbian Data" from the "Shared Task: Unsupervised MT and Very Low Resource Supervised MT" [15] with a vocabulary of 251358 words. The frequencies of the phones, diphones, and triphones were calculated and the prompts were selected from the domain-general (CV) and the domain-specific (SL) sentences, according to the scoring algorithm presented in [16] applied on diphones.

Another important criterion is the estimated duration of spoken sentences and our target was to achieve at least 20 minutes recordings per participant. To estimate the phoneme and word rate in spoken utterance we used two different approaches. According to [17] the phoneme rate was expected to be in a range of 10 (reading poetry) to 15 (commenting sport) phonemes per second. The second one was based on the word rate estimating the values in a range from 100 (English BBC) to 160 (Spanish RNE) words per minute [18]. Using these figures, we were able to estimate the expected range of speech duration by the word and phoneme counts in the selected sentences. From both HSB and CV data, we generated disjointed prompt sets, to be recorded by different speakers, HSB-1, HSB-2, and HSB-3 (in total 1200 prompts).

#### 4.2.2 Speech Recordings

The prerequisite for good acoustic model adaptation is collecting recordings from gender and age balanced groups of speakers. In total 30 speakers were recruited: 10 females, 10 males, and 10 children. 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.

Each speaker in the recording session was instructed to read the prompts exactly as displayed. Therefore it was possible to target specific pronunciation variants, to reduce the effort for post-processing and manual transcription of the recordings. The order of the prompts was randomized, and the speakers were free to read as many as they like.

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. We collected speech recordings in duration of 11 hours and 30 minutes achieving almost equal distribution across the three datasets. Around 2/3 of the recordings are considered to be used for acoustic modeling and 1/3 for testing of the "Smart Lamp" application.

## 5 Cross-lingual Acoustic Modeling

### 5.1 Forced-alignment Experiments

Using knowledge-based phoneme mapping, phoneme recognition performance is evaluated on the HSB corpus to investigate the most frequent phoneme recognition confusions.

After the initial evaluation of the confusion matrices, it was evident that there are many confusions across the vowels (i.e. /a/ with /a:/, /aI/, /aU/ ...). Therefore, the first data-driven based optimization was to reduce the phoneme set of the baseline German acoustic model (43) to only those (29) defined in Table 1. This narrows the choice of phoneme sequences and improves the robustness of the model and provides fewer confusions along with improved phoneme recognition accuracy. 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).

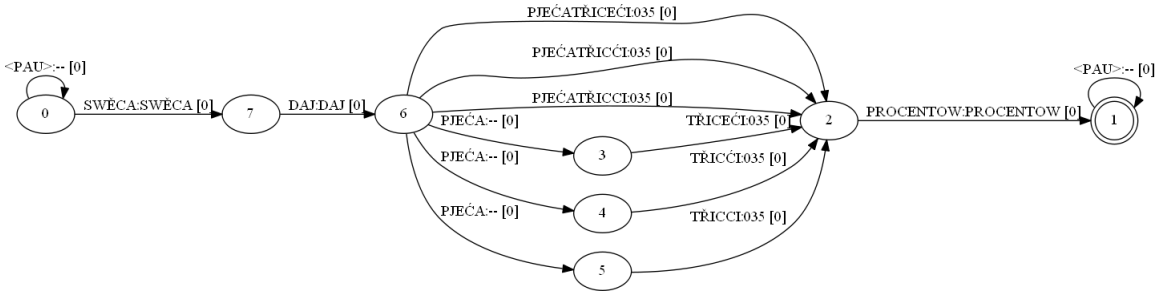


**Figure 1** – The FSG rule matching the utterance "Swěca daj pjećatřiceći procentow".

### 5.2 Acoustic Model Adaptation

The adaptation and evaluation of the acoustic models was performed with "Leave One Group Out" cross-validation (LOGO). Two of the HSB sets were used for model adaptation and tested on the third set. For instance, adaptation on the baseline and reduced models with HSB-1 and HSB-2 and their evaluation on HSB-3, etc. The results are aggregated into one data frame where for each recognized sentence, the speaker recordings were not part of the model adaptation. 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 evaluation parameters, correctness (Cor) and error (Err) were calculated over a sequence alignment using Levenstein distance. The label density (LD) indicates the ratio in the phoneme's counts of reference and the phoneme recognition. It is evident that the reduction of the phoneme set improved the correctness and reduced the errors.

**Table 3** – Phoneme recognition evaluation

| Model | Adaptation Data | Cor (%) | Err (%) | LD (%) |
|---|---|---|---|---|
| baseline | none | 42.9 | 66.9 | 99.8 |
| reduced | none | 49.4 | 61.7 | 100.1 |
| baseline | HSB (LOGO) | 66.6 | 40.0 | 96.4 |
| reduced | HSB (LOGO) | 68.9 | 37.9 | 96.3 |

From Table 3 we can see the absolute improvement in error rates (Err) after adaptation: from 66.86% to 39.98% for the baseline acoustic model, and from 61.7% to 37.9% for the reduced phoneme set model.

## 6 Smart Lamp Voice Application

The language model for the Smart Lamp application was written in the form of a set of Finite-State-Grammar (FSG) rules. To identify errors and problematic rules we tested and optimized the grammar on the adapted acoustic models to ensure the best recognition performance. The optimization was mostly directed to discover pronunciation variants and speech rate issues with compound words. Additionally, we introduced semantic tags for the intents and entities (Figure 1). After optimizing the grammar, the performance was evaluated in terms of word recognition error (Err) rate (Table 4). We achieved best recognition error of 6.7% in the case of reduced and LOGO adapted models. However, to deliver a demonstrator that will be both, speaker-independent and robust against varying acoustic environments, we adapted the reduced model exclusively on the Common Voice speech data and tested on the HSB data.

The increase in the error rate (8.7%) was not drastic and it is expected that the model will perform reliably according to the quality of the audio signal.

**Table 4** – Word recognition evaluation (Smart Lamp FSG)

| Model | Adaptation Data | Cor (%) | Err (%) | LD (%) |
|---|---|---|---|---|
| baseline | HSB (LOGO) | 93.9 | 7.2 | 99.5 |
| baseline | Common Voice | 91.0 | 10.1 | 98.5 |
| reduced | HSB (LOGO) | 94.4 | 6.7 | 100.0 |
| reduced | Common Voice | 92.4 | 8.7 | 99.2 |

# 7 Conclusions

The preliminary study confirms the feasibility of 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 to 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 a simple command and control (C&C) voice application (Smart Lamp). However, in this early stage, the used technology imposes some limitations. To improve the robustness and maximize the recognition performance, the system should be used as speaker-dependent, and the acoustic models adapted to a speaker and the acoustic environment (microphone, audio-interface, room, background noise, etc.). Each user would have a speaker profile usually created within an 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, dictation).

# 8 Acknowledgments

This study was supported by the Foundation for the Sorbian People in Bautzen, Germany. The authors would like to thank Mr. Jan Budar, Mrs. Michaela Moosche, Mrs. Katharina Čornak, and Mr. Christian Richter for their support and assistance.

# References

[1] KINSELLA, B.: *Nearly 90 million us adults have smart speakers, adoption now exceed one-third of consumers. Voicebot. URL: https://voicebot.ai*, 4, p. 28, 2020. Accessed: 2021-01-05.

[2] TEMPLETON, G.: *Language support in voice assistants compared (2019 update)*. `https://www.globalme.net/blog/languagesupport-voice-assistants-compared#Alexas_Language_Support`, 2020. Accessed: 2020-12-05.

[3] GEORGILA, K., A. LEUSKI, V. YANOV, and D. TRAUM: *Evaluation of off-the-shelf speech recognizers across diverse dialogue domains*. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6469–6476. European Language Resources Association, Marseille, France, 2020. URL `https://www.aclweb.org/anthology/2020.lrec-1.797`.

[4] BESACIER, L., E. BARNARD, A. KARPOV, and T. SCHULTZ: *Automatic speech recognition for under-resourced languages: A survey. Speech Communication, January 2014*, 56, pp. 85–100, 2014.

[5] SCHULTZ, T. and A. WAIBEL: *Language-independent and language-adaptive acoustic modeling for speech recognition*. *Speech Communication*, 35(1-2), pp. 31–51, 2001.

[6] JODLBAUER, R.: *Die aktuelle Situation der niedersorbischen Sprache : Ergebnisse einer soziolinguistischen Untersuchung der Jahre 1993–1995*. 2001.

[7] MARTI, R., B. ANDREEVA, and W. BARRY: *GENIE: The corpus for spoken lower sorbian (GEsprochenes NIEdersorbisch). The Phonetician*, 101/102, pp. 47–59, 2010.

[8] BARTELS, H. and K. THORQUINDT-STUMPF: *Ein neues Ton-und Textarchiv des muttersprachlich-dialektalen Niedersorbischen*. 2013.

[9] SCHUSTER-ŠEWC, H.: *Grammar of Upper Sorbian language: phonology and morphology*, vol. 3 of *LINCOM Studies in Slavic Linguistics*. Lincom Europa, München, 1996.

[10] HOWSON, P.: *Upper Sorbian*. *Journal of the International Phonetic Association*, 47(3), pp. 359–367, 2017.

[11] HOFFMANN, R., M. EICHNER, and M. WOLFF: *Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system*. In A. ESPOSITO, M. FAUNDEZ-ZANUY, E. KELLER, and M. MARINARO (eds.), *International Workshop on Verbal and Nonverbal Communication Behaviours. COST Action 2102*, vol. 4775 of *Lecture Notes in Computer Science*, pp. 200–218. Springer-Verlag, Vietri sul Mare, Italy, 2007. ISBN 978-3-540-76441-0.

[12] *UASR phoneme sets*. `https://rawgit.com/matthias-wolff/UASR/master/manual/index.html?reference/UasrPhonemeSets.html`, 2007. Accessed: 2021-01-05.

[13] ŠOŁĆINA, J. and E. WORNAR: *"Obersorbisch im Selbststudium/Hornjoserbšćina za samostudij"*, vol. 3 of *Ein Sprachkurs für Unerschrockene (inkl. CD)*. Aufl., Bautzen: Domowina-Verlag, 2012.

[14] ARDILA, R., M. BRANSON, K. DAVIS, M. HENRETTY, M. KOHLER, J. MEYER, R. MORAIS, L. SAUNDERS, F. M. TYERS, and G. WEBER: *Common Voice: A massively-multilingual speech corpus*. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4211–4215. 2020.

[15] BARRAULT, L., O. BOJAR, F. BOUGARES, R. CHATTERJEE, M. R. COSTA-JUSSÀ, C. FEDERMANN, M. FISHEL, A. FRASER, Y. GRAHAM, P. GUZMAN, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, A. MARTINS, M. MORISHITA, C. MONZ, M. NAGATA, T. NAKAZAWA, and M. NEGRI (eds.): *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, Online, 2020. URL `https://www.aclweb.org/anthology/2020.wmt-1`.

[16] BERRY, J., L. FADIGA ET AL.: *Data-driven design of a sentence list for an articulatory speech corpus.* In *INTERSPEECH*, pp. 1287–1291. 2013.

[17] FONAGY, I. and K. MAGDICS: *Speed of utterance in phrases of different lengths*. *Language and Speech*, 3(4), pp. 179–192, 1960.

[18] RODERO, E.: *A comparative analysis of speech rate and perception in radio bulletins*. *Text & Talk*, 32(3), pp. 391–411, 2012.