A Corpus of Read and Spontaneous Upper Saxon German Speech for ASR Evaluation

Robert Herms[1], **Laura Seelig**[1], **Stefanie Münch**[2], **and Maximilian Eibl**[1]
[1]Chair Media Informatics, Technische Universität Chemnitz, Germany
[2]FIITS, University of Applied Sciences Dresden, Germany
{robert.herms, maximilian.eibl}@cs.tu-chemnitz.de

Abstract

In this Paper we present a corpus named SXUCorpus which contains read and spontaneous speech of the Upper Saxon German dialect. The data has been collected from eight archives of local television stations located in the Free State of Saxony. The recordings include broadcasted topics of news, economy, weather, sport, and documentation from the years 1992 to 1996 and have been manually transcribed and labeled. In the paper, we report the methodology of collecting and processing analog audiovisual material, constructing the corpus and describe the properties of the data. In its current version, the corpus is available to the scientific community and is designed for automatic speech recognition (ASR) evaluation with a development set and a test set. We performed ASR experiments with the open-source framework sphinx-4 including a configuration for Standard German on the dataset. Additionally, we show the influence of acoustic model and language model adaptation by the utilization of the development set.

Keywords: speech corpus, speech recognition, dialect

1. Introduction

The collection of regional speech patterns and its preparation is the basis for many fields in linguistic research as well as the development and evaluation of speech technology applications. The dialect Upper Saxon is attributed to the East Central German dialect groups. It is mainly spoken in the Free State of Saxony and adjacent areas as the eastern part of Thuringia and the south-eastern Saxony-Anhalt. Upper Saxon has major phonological, morphological and lexical deviations from Standard German and other regiolects (Weise and others, 2013). Furthermore, there are regional variants within Saxony (Schaufuß, 2015) which arise particularly in spontaneous speech.

Some recordings of the dialect are already available by different German speech corpora. For instance, the corpus RVG 1 (Burger and Schiel, 1998) includes recordings from all over Germany. It contains read and spontaneous speech that were orthographically and phonetically transcribed. The corpus of FOLK (Schmidt, 2014) contains over 100 hours of different verbal interaction types from different regions including Upper Saxon. In other works, Upper Saxon speech has been collected in connection with specific research activities, such as the intonation of a dialect (Kügler, 2003), or standard-dialect variation (Schaufuß, 2015).

The project "Pilotprojekt zur Digitalisierung der Senderarchive sächsischer Lokalfernsehsender" in cooperation with the SLM (Sächsische Landesanstalt für privaten Rundfunk und neue Medien) tries to support local television stations by developing solutions for archiving analog audiovisual material from the early nineties. In this context, we built a comprehensive workflow including the digitization and annotation in order to provide a holistic approach. As part of the workflow, the automatic speech recognition (ASR) allows the retrieval of spoken language. However, the challenge is to recognize dialect speech.

In this Paper we present the first version of the SXUCorpus which includes read and spontaneous speech of the Upper Saxon German dialect. The main goal of this work is to encourage scientists in the field of speech technologies to handle the diversity of spoken languages. For instance, the corpus can be utilized in combination with other datasets in order to perform dialect detection or classification. Moreover, the investigation of the impact on ASR systems could be helpful for optimization problems. Therefore, our aim was to construct a corpus which has the following characteristics:

- Pure Upper Saxon dialect speech of semi-professional TV presenters and interview partners
- Regional variation of the recordings in Saxony
- Read and spontaneous speech
- Diversity of topics including news, economy, weather, sport, and documentation
- Availability to the scientific community as ASR evaluation data including a development set and a test set

This Paper is organized as follows: In the next section we present the sources and describe the data of our constructed corpus. In Section 3 we introduce the first ASR evaluation of the corpus including the experimental setup and the results as a baseline. Finally, we conclude this paper in Section 4 and give some future directions.

2. Corpus

2.1. Collecting the Data

In cooperation with the SLM we collected 289 videotapes from the archives of local television stations situated in the Free State of Saxony. These tapes comprise the analog formats SVHS and Betacam SP with a duration of up to 240 minutes. The content includes broadcasted topics of news, economy, weather, sport, and documentation from the years

Television station	# sentences		# tokens	# minutes
	read	spontaneous		
eff3	229	106	6,045	45
Nordsachsen TV	401	49	7,577	55
Lokalstudio Bischofswerda	81	65	2,077	15
VRF	103	108	4,514	31
Laubuscher Heimatkanal	0	223	4,571	33
Elsterwerda TV	0	99	2,359	18
MEF	479	744	25,817	199
Sachsen Fernsehen	353	437	17,008	104
Total	1,646	1,831	69,968	500

Table 1: Amount of data collected from different television stations in Saxony

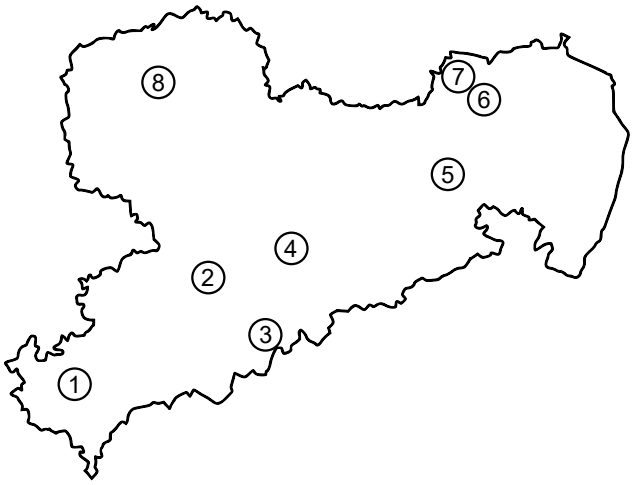

Figure 1: Regions of local television stations in the Free State of Saxony. The corresponding cities are assigned in brackets: 1 – VRF (Plauen); 2 – Sachsen Fernsehen (Chemnitz); 3 – MEF (Marienberg); 4 – eff3 (Freiberg); 5 – Lokalstudio Bischofswerda (Bischofswerda); 6 – Elsterwerda TV (Hoyerswerda); 7 – Laubuscher Heimatkanal (Höhenbocka); 8 – Nordsachsen TV (Eilenburg).

1992 to 1996. The speakers are mostly semi-professional TV presenters and interview partners. The local television stations and the corresponding cities of Saxony are illustrated in Figure 1. The stations are widely scattered all over Saxony which enables us to derive benefit from a high variety of the Upper Saxon speech. A regional aggregation would not reproduce the diversity of the dialect.

In order to manage the digitization of the analog formats we developed a robot-system and a software framework called Imtecs (Manthey et al., 2013). This framework provides a collection of components to implement an automated ingest workflow (as part of a holistic archive workflow) which in addition enables the extraction of metadata of the technical constraints for each ingested job. The resulting raw digital audiovisual formats are then transcoded into WAV (mono, 16 kHz, 16 bit) to obtain an appropriate format for the annotation described in the next section.

2.2. Corpus Construction

The transcription of the acquired data was conducted by four student assistants who are trained transcribers and familiar with the occurring dialect. We used the software FOLKER (Schmidt and Schütte, 2010) for transcribing and utilized the minimal variant of the GAT transcription convention (Selting et al., 2009). The segmentation was performed on sentence-level. Only segments were considered that have a low background noise and include only speech with the typical characteristics of the Upper Saxon dialect. Standard German or other dialects respectively languages were left out. In case of indecision of a transcriber we preferred to conduct a majority vote to fulfill our requirements of the corpus.

Some words in the data yields dialect specific phenomena concerning the utilization of an alternative pronunciation. Consequently different official orthographic forms of a certain word were applied. For instance, the interjections “nu”, “na”, and “nor” often appeared in spontaneous speech and are instances of an affirmation such as “ja” in standard orthography respectively represent an agreement. Another often used Upper Saxon expression which serves as an example is “bissel”. It means “a bit” in English and has the form “bisschen” in Standard German.

As there is a huge amount of different speakers and the appearance of a single speaker is rather low we decided to describe speakers by a label for gender (male and female) and the type of speech (read and spontaneous) instead of a unique speaker name or identifier. The labels and transcriptions were double checked by another student assistant guaranteeing a higher quality of the edited data.

Next, we parsed each XML based project file of FOLKER in order to extract transcribed audio segments (WAV, mono, 16 kHz, 16 bit) from a long WAV file. Additionally we assigned the labels, the ID of television station and videotape, and the number of the segment to the name of a segmented audio file. In order to generate statistics of the transcripts and to prepare the corpus concerning ASR we generated a *.transcription* file where each line represents a sentence with an assigned name of the corresponding audio segment in brackets and a *.fileid* file which includes the references to the audio segments.

The statistics of the constructed corpus are shown in Table 1. The corpus comprises in its first version approximately

Dataset	# sentences	# tokens	# minutes
read – development set	823	14,566	112
read – test set	823	14,310	112
spontaneous – development set	916	20,583	138
spontaneous – test set	915	20,516	138
both – development set	1,739	35,149	250
both – test set	1,738	34,826	250

Table 2: Amount of data prepared for ASR evaluation

Configuration	read		spontaneous		both	
	dev	test	dev	test	dev	test
Standard	52.2	51.4	77.9	77.1	67.2	66.5
AM adaptation	13.9	30.2	34.5	56.1	32.5	46.2
LM adaptation	34.8	50.9	66.8	75.2	55.4	65.1
AM + LM adaptation	6.3	29.6	21.9	53.4	20.7	44.3

Table 3: Word error rates for read, spontaneous, and both types of speech using different ASR configurations on the development set (dev) and the independent test set (test).

500 minutes and 69,968 transcribed tokens. At this stage, the dataset is not suitable for the investigation of different speech patterns (e.g., pronunciation variation) in the state of Saxony itself, since the number of sentences is unbalanced concerning the individual regions. Nevertheless, the total corpus has two things in common: the Upper Saxon dialect and a comparatively good partitioning of read and spontaneous speech. This version of the corpus is therefore prioritized for the evaluation of ASR systems.

For this purpose, we prepared the corpus as follows (see Table 2). We separated the dataset into a read and a spontaneous speech part. For each part we assigned the same number of segments originated from a videotape to the test set and to the development set. Thus, the regional distribution is balanced in the generated evaluation sets as well as for the type of speech. Finally, we created a third evaluation set without the consideration of read and spontaneous speech by merging the two development sets and the two test sets.

3. ASR Evaluation

3.1. Experimental Setup

The goal of this experiment is to verify the dataset by ASR with a configuration for Standard German, but also to investigate the influence of the development set concerning acoustic model (AM) and language model (LM) adaptation. ASR was performed using the engine of the open-source framework sphinx-4 (Walker et al., 2004). We utilized the German open source corpus for distant speech recognition (Radeck-Arneth et al., 2015) to train an AM. It includes context-dependent triphone Hidden-Markov-Models (HMM) with 2,000 senones and 32 Gaussians per state. In order to construct a LM with a high variety of topics we used approximately 15 million German sentences of newspaper texts from the Leipziger Corpus Collection (Richter et al., 2006). We trained a trigram LM with Kneser-Ney smoothing by using the SRILM toolkit (Stolcke et al., 2011). The perplexity of the LM for our total corpus is 588.1 and has an Out-of-Vocabulary (OOV) rate of 0.9%. The pronunciation lexicon was automatically created with the Balloon tool (Reichel, 2012) covering the vocabulary of the acquired newspaper texts. Subsequently, we corrected these canonical pronunciations manually in terms of Standard German.

One approach to achieve a better ASR performance is AM adaptation, e.g., (Wang et al., 2003). We applied maximum a posteriori (MAP) adaptation by using the development set of read, spontaneous, and both to update the parameters of our acoustic model.

In order to consider the characteristics of the language especially in spontaneous speech we performed LM adaptation as follows. We used the transcripts of the development set of read, spontaneous, and both to train the corresponding trigram LMs with Kneser-Ney smoothing. For each resulting model we performed a linear interpolation with the background model constructed with the newspaper texts from the Leipziger Corpus Collection. For the adaptation, equal interpolation weights were assigned.

3.2. Results

We report different word error rates (WER) obtained on the development set and the independent test set with different configurations of ASR in Table 3. As expected, there are much more recognition errors in spontaneous (77.1%) than in read speech (51.4%) with a difference of 25.7%. AM adaptation using the development set of the corresponding type of speech yields better results than the standard configuration. In more detail, we achieved an improvement of 21.2% for read, 21.0% for spontaneous, and 20.3% for both types of speech. With our background LM and the test set we computed the perplexities 548.6 (read), 644.1 (spontaneous), and 602.8 (both). After LM adaptation, the perplexities could be decreased to 497.7 (read), 411.9 (spontaneous), and 443.3 (both). However, the WER results with LM adaptation were just slightly better than using only the background model. We obtained the best results on the

test set by the combination of AM and LM adaptation with 29.6% (read), 53.4% (spontaneous), and 44.3% (both). One can obtain a decrease in WER by more sophisticated models as well as a detailed investigation regarding pronunciation variation.

4. Conclusions and Future Work

We presented the SXUCorpus comprising read and spontaneous speech of the Upper Saxon German dialect. The data has been collected from archives of local television stations located in the Free State of Saxony. This paper reports the methodology of collecting and processing the analog audiovisual material, constructing the corpus and describes the properties of the data. In its current version, the corpus is available on request to the scientific community and is designed for ASR evaluation including a development set and a test set. We performed ASR experiments with a configuration for Standard German on the dataset. Additionally, we showed the effect of acoustic model and language model adaptation. We could decrease the word error rate, with a difference of more than 20%, to 29.6% for read, 53.4% for spontaneous, and 44.3% for both types of speech.

For the future we plan to set further projects in order to verify and enhance our archiving solutions. In this connection, we want acquire more and more speech of the Upper Saxon dialect. More speech data and more balancing data enable the investigation of pronunciation variation as well as different regional speech patterns.

5. Acknowledgements

This work was supported by the SLM (Sächsische Landesanstalt für privaten Rundfunk und neue Medien).

6. References

Burger, S. and Schiel, F. (1998). Rvg 1-a database for regional variants of contemporary german. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 1083–1087.

Kügler, F. (2003). Do we know the answer? variation in yes-no-question intonation. *Linguistics in Potsdam*, 21:9–29.

Manthey, R., Herms, R., Ritter, M., Storz, M., and Eibl, M. (2013). A support framework for automated video and multimedia workflows for production and archive. In *Human Interface and the Management of Information. Information and Interaction for Learning, Culture, Collaboration and Business*, pages 336–341. Springer.

Radeck-Arneth, S., Milde, B., Lange, A., Gouvêa, E., Radomski, S., Mühlhäuser, M., and Biemann, C. (2015). Open source german distant speech recognition: Corpus and acoustic model. In *Text, Speech, and Dialogue*, pages 480–488. Springer.

Reichel, U. D. (2012). Perma and balloon: Tools for string alignment and text processing. In *INTERSPEECH*, pages 1874–1877. Citeseer.

Richter, M., Quasthoff, U., Hallsteinsdóttir, E., and Biemann, C. (2006). Exploiting the leipzig corpora collection. *Proceesings of the IS-LTC*.

Schaufuß, A. (2015). Standard-dialect variation and its functionalization. In *Language Variation-European Perspectives V: Selected papers from the Seventh International Conference on Language Variation in Europe (ICLaVE 7), Trondheim, June 2013*, volume 17, page 183. John Benjamins Publishing Company.

Schmidt, T. and Schütte, W. (2010). Folker: An annotation tool for efficient transcription of natural, multi-party interaction. In *LREC*.

Schmidt, T. (2014). The research and teaching corpus of spoken german-folk. In *LREC*, pages 383–387.

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J. R., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., et al. (2009). Gesprächsanalytisches transkriptionssystem 2 (gat 2). *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion*.

Stolcke, A., Zheng, J., Wang, W., and Abrash, V. (2011). Srilm at sixteen: Update and outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, page 5.

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., and Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition. Technical report, Mountain View, CA, USA.

Wang, Z., Schultz, T., and Waibel, A. (2003). Comparison of acoustic model adaptation techniques on non-native speech. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–540. IEEE.

Weise, C. et al. (2013). Upper saxon (chemnitz dialect). *Journal of the International Phonetic Association*, 43(02):231–241.