Data-Driven Pronunciation Modeling of Swiss German Dialectal Speech for Automatic Speech Recognition

Michael Stadtschnitzer, Christoph Schmidt
Fraunhofer IAIS, Schloss Birlinghoven, 53757 Sankt Augustin
{michael.stadtschnitzer, christoph.andreas.schmidt}@iais.fraunhofer.de

Abstract

Automatic speech recognition is a requested technique in many fields like automatic subtitling, dialogue systems and information retrieval systems. The training of an automatic speech recognition system is usually straight forward given a large annotated speech corpus for acoustic modeling, a phonetic lexicon, and a text corpus for the training of a language model. However, in some use cases these resources are not available. In this work, we discuss the training of a Swiss German speech recognition system. The only resources that are available is a small size audio corpus, containing the utterances of highly dialectical Swiss German speakers, annotated with a standard German transcription. The desired output of the speech recognizer is again standard German, since there is no other official or standardized way to write Swiss German. We explore strategies to cope with the mismatch between the dialectal pronunciation and the standard German annotation. A Swiss German speech recognizer is trained by adapting a standard German model, based on a Swiss German grapheme-to-phoneme conversion model, which was learned in a data-driven manner. Also, Swiss German speech recognition systems are created, with the pronunciation based on graphemes, standard German pronunciation and with a data-driven Swiss German pronunciation model. The results of the experiments are promising for this challenging task.

Keywords: Robust Speech Recognition, Swiss German, Dialectal Speech, Data-Driven Pronunciation Modeling

1. Introduction

Switzerland has four national languages: German/Swiss German (63%), French (22.7%), Italian (8.1%), Romansh (0.5%); the numbers in brackets are the percentages of the population speaking them[1]. As can be derived from Figure 1, French is spoken in the west, Italian is spoken primarily in Ticino, Val Bregaglia and Val Pschiavo and Romansh speakers are distributed over Graubünden. Swiss German, which is primarily spoken in the center and east of Switzerland, is highly dialectical. Typically, speakers speak a dialect representative of the region. To be understood by visitors, Swiss German speakers switch to standard German (Garner et al., 2014). Swiss German and its dialectal variants do not have a standard written form, instead the standard written form is standard German. Although regional Swiss German dialects are manifold and their differences can be very subtle, there is a common Swiss German speaking style which is used in Swiss broadcasts (e.g. weather reports of *Schweizer Radio und Fernsehen*[2] (SRF)), and that is understood well by the vast majority of the Swiss German speaking people. Broadcast companies are typically interested in the automatic transcription of speech data including (live) subtitling and automatic speech recognition (ASR). In the case of Swiss German, the desired output of the ASR system is standard German, as there is no standardized written form of Swiss German dialect. Due to the mismatch between the dialectal pronunciation of Swiss German and the written form of standard German, these system often fail. In this paper we are exploring methods to close this gap, amongst others a data-driven method.

In the remainder of the paper, we first describe the data investigated in this work and the available annotation (Section 2.). In Section 3., we then describe the involved ASR recognition methods including the data-driven method to improve the pronunciation model and evaluate the results. A conclusion is given in Section 4..



Figure 1: Geographical distribution of the languages of Switzerland (2000); Swiss Federal Statistical Office, `www.bfs.admin.ch`

2. Resources

This section describes the speech corpora that are used in this work.

2.1. SRF Meteo weather report dataset

In this paper we describe the Swiss German SRF Meteo dataset, which *Schweizer Radio und Fernsehen* generously provided us for research purposes. This dataset consists of Swiss German weather reports of SRF Meteo. The speakers speak Swiss German dialect, and the textual annotation is standard German. The dataset consists of 290 Meteo weather report broadcasts with a total of 10,201 speech segments and a total of 6.5 hours of annotated speech and a total of 83,449 annotated words. The contained speech is

[1] `http://www.swissinfo.ch`
[2] `http://www.srf.ch`

to a large extent about weather forecasts and contain a large number of place names.

2.2. GerTV1000h German Broadcast corpus

In LREC 2014 we presented the German broadcast corpus GerTV1000h (Stadtschnitzer et al., 2014). The corpus consists of approximately 1,005 hours of German speech data from the broadcast domain and covers a broad selection of news, interviews, talk shows and documentaries, both from television and radio across several stations. The data subsets from this corpus that were used in this work are summarized in Table 1. Note that we discarded a small amount of utterances in the training set due to mispronunciations, unintelligible words and word fragments.

2.3. Difficult Speech Corpus (DiSCo)

In LREC 2010, Fraunhofer IAIS presented the Difficult Speech Corpus (DiSCo) (Baum et al., 2010). The DiSCo corpus is a collection of datasets from German broadcast domain with challenging acoustical situations. It is intended for the evaluation of speaker and speech recognition systems. The datasets are separated into planned and spontaneous speech. Challenging acoustical situations that are covered by the data subsets are clean, music, applause, and mixed condition. In this work we use the DiSCo corpora for evaluation purposes. The DiSCo data subsets used in this work are listed in Table 2.

3. Experimental setup

3.1. Motivation

The training of a speech recognition system given annotated speech data, a pronunciation model and text is straight forward. However in this setup, we want to train a speech recognition system which is able to translate highly dialectal Swiss German speech data in standard German text. This is desirable, because there is no standardized way of writing Swiss German other than standard German.

3.2. Data preparation

For the experiments the Meteo dataset was split into a training, a development and a testing set, as can be seen in Table 3. We choose to have 260 weather reports in the training set and each 15 weather reports in the development and the testing set. The distribution of the weather reports into the datasets was performed randomly. When considering only the text of the training set for the training of a language model, the development set and the test set have an out-of-vocabulary (OOV) rate of $\text{OOV}_{\text{dev}} = 7.6\%$ and $\text{OOV}_{\text{test}} = 9.1\%$. This seems quite high, however the running OOV rate is acceptable considering the small amount of training data, namely $\text{OOV}_{\text{r,dev}} = 1.4\%$ and $\text{OOV}_{\text{r,test}} = 1.7\%$.

3.2.1. Standard German Speech Recognition

To perform the following data-driven experiments, where a standard German phoneme recognizer is involved, we need to train a standard German speech recognition system. In (Schmidt et al., 2016) we proposed the standard German broadcast ASR system based on recurrent neural networks (RNN) as implemented in (Miao et al., 2015). For the lexicon, we use the pronunciation model trained with Sequitur G2P (Bisani and Ney, 2008) and the German pronunciation database Phonolex (BAS - Bavarian Archive for Speech Signals, 2013). We use the 1-best pronunciation of the model for each word of the lexicon. We use the broadcast text corpora consisting of 75 millions of words, which we already used in other works (Stadtschnitzer et al., 2014) for the training of the language model. We use a 5-gram model, trained with modified shift beta algorithm with back-off weights using IRSTLM (Federico et al., 2008) and a dictionary size of approximately 500,000 words and a language model pruning factor of 10^{-8}. For training of the acoustical model, the GerTV1000h corpus (Section 2.2.) was used.

By the use of time delay neural network (TDNN) architecture with speed-perturbated training data as proposed in (Peddinti et al., 2015) and implemented in the Kaldi Toolkit (Povey et al., 2011), we were able to improve the speech recognition results on the DiSCo evaluation sets as indicated in Table 4. By both the employment of projected long-short memory networks (LSTMP) in the TDNN architecture as proposed in (Cheng et al., 2017) and the use of gated convolutional neural networks (GCNN) (Dauphin et al., 2017) (as implemented in TheanoLM Toolkit (Enarvi and Kurimo, 2016)) for n-best hypotheses rescoring ($n = 200$) we could further improve our standard German broadcast ASR system.

3.3. Swiss German Data-Driven Pronunciation

For the experiments regarding the data-driven Swiss German pronunciation model, we employ the TDNN ASR model (as described in Section 3.2.1., because this was the best configuration available at the time point of the experiments. The results of the standard German TDNN ASR system, which performed well for the standard German evaluation data (Table 4), are naturally worse on the Meteo data ($\text{WER}_{\text{dev}} = 81.0\%$, $\text{WER}_{\text{test}} = 79.5\%$), since there is a large mismatch in speech, phonetics and language between standard German and Swiss German. By replacing the language model trained from broadcast text by a language model trained on the text of the Meteo training dataset, we can reduce this mismatch for the Meteo evaluation data to $\text{WER}_{\text{dev}} = 64.98\%$ and $\text{WER}_{\text{test}} = 64.73\%$. In the following we try to further reduce the mismatch, especially for the mismatch caused in the pronunciation, in a data-driven manner.

3.3.1. Standard German Speech Phoneme Decoder

We first create a phoneme decoder and then use the phoneme decodings to create a Swiss German G2P model. For the training of the standard German Speech phoneme decoder, we use the TDNN acoustical models discussed in Section 3.2.1.. For the training of the standard German phoneme language model, which is required for the phoneme decoder, we replace the words from the text of the Meteo training dataset by its pronunciations derived from the standard German G2P model. Then we train a 5-gram phoneme language model and use it for decoding of the speech signals.

Dataset	#Segments	#Words	Avg. Words	#Unique	Size(h)	Avg. Length (s)
Train	773,631	9,406,119	12.2	243,313	991.9	4.6
Dev	2,348	33,748	14.4	6,377	3.5	5.3

Table 1: Statistics of the GerTV1000h data subsets used in this work

Dataset	#Segments	#Words	Avg. Words	#Unique	Size(h)	Avg. Length (s)
Planned, clean	1,364	9,184	6.7	2,939	0.9	2.4
Spontaneous, clean	2,861	20,740	7.4	4,019	1.9	2.4

Table 2: Statistics of the DiSCo data subsets used in this work

Dataset	#Shows	#Segments	#Words	Avg. Words	#Unique	Size(h)	Avg. Length (s)
Meteo train	260	9,181	75,215	8.2	2,981	5.9	2.3
Meteo dev	15	493	3,995	8.1	742	0.3	2.2
Meteo test	15	527	4,242	8.1	778	0.3	2.2

Table 3: Statistics of the Meteo data subsets used in this work

Model	GerTV dev	DiSCo planned	DiSCo spont.
RNN	17.2	11.9	14.5
TDNN	15.6	11.1	13.2
TDNN-LSTMP	13.7	8.9	10.4
TDNN-LSTMP-GCNN	**12.7**	**8.1**	**9.3**

Table 4: WER [%] results of the standard German speech recognition systems

3.3.2. Data-Driven Pronunciation Modeling

By decoding the Swiss German Meteo training set using the phoneme language model, we get some suggestions of how the speech in the audio data was pronounced. However, the data is organized in utterances, rather than words. Nonetheless, we train a Swiss German G2P model by using the phrases (whitespaces are replaced by an underscore) followed by the pronunciations from the phoneme decodings. The trained Swiss German pronunciation model is able to provide some good suggestions in the n-best list for the pronunciation of several words, as can be derived from Table 5. In this table we also show a non-standardized Swiss German dialectal text annotation of an online Swiss German dictionary[3] for comparison. The pronunciations from the Swiss German G2P were found in a data-driven manner, without any knowledge of the online Swiss German dictionary. As can be learned from Table 5 the pronunciations learned from the Swiss German G2P are often quite near to the textual correspondents from the online Swiss German dictionary.

We then created several lexicons, which were composed by the 1-best standard German pronunciation and a n-best list of the data-driven Swiss German G2P. The intention was to keep the 1-best standard German pronunciation as a backup, when no meaningful Swiss German pronunciation can be found by the method. We then used the enriched lexicons with the standard German TDNN models and a language model trained on the text from the Meteo training dataset. The results are depicted in Figure 2. We optimized the parameter n on the Meteo development set. Derived from the results of the experiments, the optimal variant is to add a 2-best list of the data-driven Swiss German G2P to the 1-best standard German pronunciations. Using this adapted configuration, which includes both reasonable Swiss German and standard German pronunciations, the WER could be reduced for the Meteo development and test set to $\text{WER}_{\text{dev}} = 60.3\%$ and $\text{WER}_{\text{test}} = 56.4\%$.

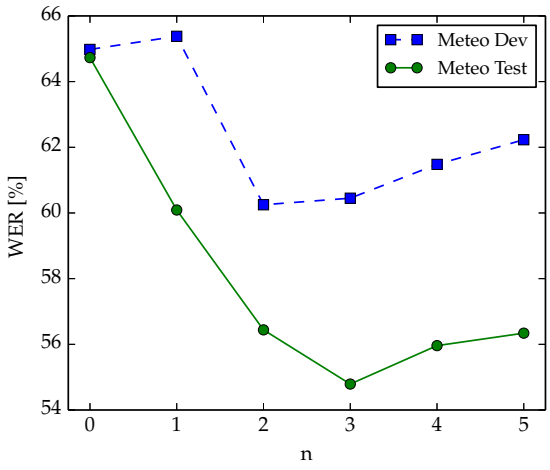


Figure 2: WER for different n for configurations with 1-best standard German pronunciation and n-best Swiss German pronunciations from the speech data driven G2P model

3.4. Supplementary Experiments

We also wanted to evaluate how far we can get, when we train the Swiss German models in a straight forward manner by either using grapheme pronunciations, standard German phoneme pronunciations or the combined pronunciation as described in Section 3.3.2.. When using a grapheme pronunciation, each word is modeled by a sequence of its graphemes (i.e. Montag ⇒ m o n t a g). When using standard German phoneme pronunciations, we use the standard German pronunciation model, which is trained on the standard German Phonolex pronunciation lexicon (BAS - Bavarian Archive for Speech Signals, 2013) using Sequitur G2P (Bisani and Ney, 2008). For the training of the acoustical model we use the training dataset of the SRF Meteo

[3] https://www.pauker.at/pauker/DE_DE/SC/wb

Standard German	German G2P	Data-Driven Swiss German G2P	Swiss German Online Dictionary
Montag	m o: n t a: k	m a: n t i: k	Mäntig
Dienstag	d i: n s t a: k	ts i: S t i: k	Ziischtig
Mittwoch	m I t v O x	m I t b u: x	Mittwuch
Donnerstag	d O n 6 s t a: k	d a n S t i: k	Danschtig
Freitag	f r aI t a: k	f r i: t I k	Fritig
Samstag	z a m s t a: k	Q a m S t i: k	Samschtig
Sonntag	z O n t a: k	z o d I k	Sunntig

Table 5: Phoneme translations of standard German words using the standard German and the speech data-driven Swiss German G2P

dataset. For training the language model we use IRSTLM toolkit (Federico et al., 2008) and we use a 5-gram model with modified shift beta algorithm with back-off weights. For training of the Swiss German ASR system, we either use Eesen (Miao et al., 2015) toolkit, when using long short term memory (LSTM) recurrent neural networks (RNN) with connectionist temporal classification (CTC) training, or the Kaldi toolkit (Povey et al., 2011), when using Hidden Markov Models with Gaussian Mixture Models (HMM-GMM), or hybrid HMM with feed forward Deep Neural Networks (HMM-DNN) or the state-of-the-art time delay neural networks with projected long short-term memory (TDNN-LSTMP) layers. The results are shown in Table 6. The HMM-GMM, DNN and TDNN-LSTMP models from the Kaldi toolkit are trained with bootstrapping and provide more stable results in this setup (i.e. a setup with a small amount of training data) compared to the RNN models, which use Connectionist Temporal Classification (CTC) instead, and which are trained directly on the audio data. It is also remarkable that there is no big difference when comparing standard German grapheme pronunciations to standard German phoneme pronunciations. Both setups perform almost equally well. The use of the combined standard German and Swiss German pronunciation performed slightly worse compared to standard German and grapheme pronunciations for the HMM-GMM case. We believe this is the case because during training the algorithm needs a consistent single pronunciation, so the algorithm can model the pronunciation including the possible mismatches consistently. The TDNN-LSTM models trained with the standard German G2P pronunciations performed best on the Meteo test set ($\text{WER}_{\text{test}} = 23.8\%$) given the experiments performed.

Model	Pron.	Meteo dev	Meteo test
HMM-GMM	GG2P	39.7	28.9
HMM-GMM	Graph.	40.3	29.6
HMM-GMM	SGG2P	41.3	30.8
RNN	GG2P	44.5	32.7
RNN	Graph.	45.0	32.3
HMM-DNN	GG2P	37.1	27.1
HMM-DNN	Graph.	37.7	27.0
TDNN-LSTMP	GG2P	34.9	**23.8**
TDNN-LSTMP	Graph.	**34.8**	24.3

Table 6: WER [%] results of directly trained Swiss German speech recognitions systems using different types of pronunciation lexicons; standard German G2P (GG2P), combined data-driven Swiss German and standard German G2P (SGG2P) or grapheme sequences (Graph.)

4. Conclusion

In this paper we explored the creation of a Swiss German speech recognition system by employing a small Swiss German dataset. Since there is no standardized way to write Swiss German other than standard German, the annotations of the Swiss German audio corpus are standard German, in contrast to the audio material which is highly dialectical Swiss German. The desired output of the Swiss German speech recognition system is again standard German. Unfortunately we lack a Swiss German pronunciation lexicon that maps standard German words into Swiss German pronunciations. We approach this problem by successfully adapting a high-performance standard German speech recognition system to the Swiss German pronunciations by the employment of a Swiss German G2P model which was learned in a data-driven manner by phoneme decodings derived from the standard German speech recognition system with the use of a phoneme language model. It turned out that by adding a 2-best list of Swiss German pronunciations derived from the data-driven Swiss German G2P model to the 1-best standard pronunciations, the adapted model provided the best results, when adapting the standard German model. However, the training of an ASR system directly on the Swiss German data by replacing the missing Swiss German pronunciation by either a standard German phoneme or grapheme sequences, provided even better results. The use of the combined lexicon did not prove to be beneficial when training a system directly on the Swiss German audio data in contrast to the adaptation of the standard German model. For both standard German and Swiss German models, the use of TDNN-LSTMP provided the best results with word error rates as low as 8.1% and 23.8% respectively on the corresponding test sets. This are encouraging results given the small amount of available training data in the Swiss German case.

5. Acknowledgements

The authors would like to thank the *Schweizer Radio und Fernsehen* (SRF) for supporting our research and for their generosity of providing us data for research purposes.

6. Bibliographical References

Bisani, M. and Ney, H. (2008). Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech Communication*, 50:434–451, July.

Cheng, G., Peddinti, V., Povey, D., Manohar, V., Khudanpur, S., and Yan1, Y. (2017). An exploration of dropout with LSTMs. In *Proceedings of INTERSPEECH*, Stockholm, Sweden, Aug.

Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. (2017). Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia.

Enarvi, S. and Kurimo, M. (2016). TheanoLM - An Extensible Toolkit for Neural Network Language Modeling. In *Proc. of INTERSPEECH*, San Francisco, USA.

Federico, M., Bertoldi, N., and Cettolo, M. (2008). IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proceedings of Interspeech*, Brisbane,Australia.

Garner, P. N., Imseng, D., and Meyer, T. (2014). Automatic Speech Recognition and Translation of a Swiss German Dialect: Walliserdeutsch. In *Proceedings of Interspeech*, Singapore, China, September.

Miao, Y., Gowayyed, M., and Metze, F. (2015). EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 167–174, Scottsdale, Arizona, USA, December.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proceedings of INTERSPEECH*, Dresden, Germany, September.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Schmidt, C., Stadtschnitzer, M., and Köhler, J. (2016). The Fraunhofer IAIS Audio Mining System: Current State and Future Directions. In *Proceedings of ITG Fachtagung*, Paderborn, Germany.

7. Language Resource References

BAS - Bavarian Archive for Speech Signals. (2013). *Pronunciation lexicon PHONOLEX*. `https://www.phonetik.uni-muenchen.de/forschung/Bas/BasPHONOLEXeng.html`.

Doris Baum and Daniel Schneider and Rolf Bardeli and Jochen Schwenninger and Barbara Samlowski and Thomas Winkler and Joachim Köhler. (2010). *DiSCo - A German Evaluation Corpus for Challenging Problems in the Broadcast Domain*. In Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC), Valetta, Malta.

Michael Stadtschnitzer and Jochen Schwenninger and Daniel Stein and Joachim Koehler. (2014). *Exploiting the large-scale German Broadcast Corpus to boost the Fraunhofer IAIS Speech Recognition System*. Proceedings of the 9th Language Resources and Evaluation Conference (LREC), 2014, Reykjavik, Iceland.