Adaptation and Training of a Swiss German Speech Recognition System using Data-driven Pronunciation Modelling

Michael Stadtschnitzer[1], Christoph Schmidt[2]

[1] *Fraunhofer IAIS, Schloß Birlinghoven, 53757 Sankt Augustin, Deutschland, michael.stadtschnitzer@iais.fraunhofer.de*

[2] *Fraunhofer IAIS, Schloß Birlinghoven, 53757 Sankt Augustin, Deutschland, christoph.andreas.schmidt@iais.fraunhofer.de*

Abstract

Automatic speech recognition is a very important technique for numerous applications like automatic subtitling, dialogue systems and information retrieval systems. Given an annotated speech corpus, a phonetic lexicon and a text corpus, the training of speech recognition systems is straight forward. However, sometimes some of these resources are not available, and strategies must be explored to fill this gap. In this work we train a Swiss German speech recognition system. The only resources that are available is a small Swiss German speech corpus, which is annotated with a standard German transcription. Standard German is the desired output of the speech recognition system, since there is no standardised way to write Swiss German. We use a data-driven approach to estimate the Swiss German pronunciations from a standard German speech recognition model to improve the Swiss German speech recognition system. Evaluations of the Swiss German speech recognition system show promising results.

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]. 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 [6]. Swiss German and its dialectal variants do not have a standard written form, instead the standard written form is standard German. The mismatch, especially in pronunciation, between spoken Swiss German and written standard German can be problematic when training speech recognition systems, especially when no machine readable pronunciation lexicon is available. That is why in this paper we are trying to approach this issue by automatically modelling the Swiss German pronunciation with a data-driven approach.

SRF Meteo Weather Forecast Dataset

In this section the SRF Meteo weather forecast dataset is described, which *Schweizer Radio und Fernsehen* (SRF) generously provided us for research purposes. It consists of 290 SRF Meteo weather forecast shows. The speakers speak Swiss German and the textual annotation is standard German. The statistics of the dataset, and the division into training, development and testing dataset are listed in Table 1. The distribution of the weather forecasts into the datasets was performed randomly.

Dataset	#Shows	#Segments	#Words	Size(h)
Full	290	10,201	83,449	6.5
Train	260	9,181	75,215	5.9
Dev	15	493	3,995	0.3
Test	15	527	4,242	0.3

Table 1: Statistics of the SRF Meteo Dataset

Standard German Speech Recognition

In this section the standard German broadcast speech recognition system, which is employed in the Fraunhofer IAIS audio mining system [10], and which is used in this work is described in this section. The current speech recognition systems, which are trained on the full GER-TV corpus [11] and evaluated on the DiSCo corpus [2], are listed in Table 2. Speed perturbation was employed to extend the training data size (i.e. 3×992 hours) in some configurations. For the training of the 5-gram (and the GCNN [5]) language model a German text corpus with 75 millions of words was used. The grapheme-to-phoneme model (G2P) [3] to derive the pronunciation lexicon is trained on Phonolex [1].

Configuration	Size (hours)	Ger-TV dev	DiSCo planned	DiSCo spont.
RNN [7]	992	17.2	11.9	14.5
TDNN [8, 9]	3x992	15.6	11.1	13.2
TDNN-LSTMP [4, 9]	3x992	13.7	8.9	10.3
TDNN-LSTMP-GCNN [5, 9]	3x992	12.7	8.1	9.3

Table 2: Word error rate (WER) [%] results of different configurations of the Fraunhofer IAIS speech recognition system

Swiss German Speech Recognition

For the adaptation of the speech recognition system, we employ the time-delay neural network (TDNN) system, as described in Table 2, which was the best configuration at the time of the experiments. On the Swiss German SRF Meteo data this configuration naturally shows degraded performance ($\text{WER}_{\text{dev}} = 81.0\%$, $\text{WER}_{\text{test}} = 79.5\%$). Using the text of the Meteo training corpus for language modelling (5-gram), the performance on the Meteo datasets can be increased to $\text{WER}_{\text{dev}} = 65.0\%$ and $\text{WER}_{\text{test}} = 64.7\%$. To retrieve phoneme decodings, we employ a 5-gram phoneme language model, derived from the Meteo training text which is translated into phoneme sequences by the standard German G2P model. From the

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

phoneme decodings of the Meteo training set, which now contain Swiss German pronunciations (using the same phoneme set), we train a Swiss German G2P model, by inputing the whole phrases with the corresponding phoneme decodings. For words that occur frequently in the audio data, the pronunciation was modelled very well. To retrieve an adapted Swiss German speech recognition model from the standard German model, we add an n-best pronunciation (data-driven Swiss German G2P) to the 1-best pronunciation list of the standard German G2P model, which is kept as a fallback pronunciation. We optimised the factor n over the Meteo development set, and found out that $n = 2$ performed best on our data. Using this method, the performance of the adapted Swiss German speech recognition system was $\text{WER}_{\text{dev}} = 60.3\%$ and $\text{WER}_{\text{test}} = 56.4\%$. For comparison, we also trained state-of-the-art speech recognition systems directly on the Meteo training dataset, by using a standard German pronunciation. The results are listed in Table 3, and favor the direct training on the Swiss German data.

Model	Meteo dev	Meteo test
RNN [7]	44.5	32.7
TDNN-LSTMP [4, 9]	34.9	23.8

Table 3: WER [%] results of directly trained Swiss German speech recognition systems

Conclusion and Outlook

In this work we adapted a standard German speech recognition system to Swiss German speech. We first described the standard German speech recognition system and the used resources. Then we replaced the standard German language model by a language model training on the text of the Swiss German training corpus, which improved the results. We then employed a phoneme decoder, derived from the standard German speech recognition system, to retrieve phoneme decodings on the Swiss German training dataset. From the phoneme decodings a Swiss German G2P was trained which is able to output Swiss German pronunciations. We then further adapted the standard German speech recognition system, by adding a 2-best list of Swiss German pronunciations to the pronunciation lexicon, which again improved the performance. However, direct training of a speech recognition system using Swiss German speech data and standard German pronunciation, where the Swiss German pronunciation modelling is performed implicitly by the acoustic model, performed better. Nonetheless, we were able to successfully adapt a standard German speech recognition system to Swiss German speech by data-driven pronunciation modelling, and were able to retrieve meaningful Swiss German pronunciations automatically.

Acknowledgements

The authors want to thank *Schweizer Rundfunk und Fernsehen* (SRF) for providing the data and assisting with the research. Without their help this research would not have been possible.

References

[1] BAS - Bavarian Archive for Speech Signals. Pronunciation lexicon PHONOLEX, 2013.

[2] Doris Baum, Daniel Schneider, Rolf Bardeli, Jochen Schwenninger, Barbara Samlowski, Thomas Winkler, and Joachim Köhler. DiSCo - A German Evaluation Corpus for Challenging Problems in the Broadcast Domain, 2010.

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

[4] Gaofeng Cheng, Vijayaditya Peddinti, Daniel Povey, Vimal Manohar, Sanjeev Khudanpur, and Yonghong Yan1. An exploration of dropout with LSTMs. In *Proceedings of INTERSPEECH*, Stockholm, Sweden, Aug 2017.

[5] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017.

[6] Philip N. Garner, David Imseng, and Thomas Meyer. Automatic Speech Recognition and Translation of a Swiss German Dialect: Walliserdeutsch. In *Proceedings of Interspeech*, Singapore, China, September 2014.

[7] Yajie Miao, Mohammad Gowayyed, and Florian Metze. 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 2015.

[8] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proceedings of INTERSPEECH*, Dresden, Germany, September 2015.

[9] Daniel Povey, Arnab Ghoshal, and Boulianne et al. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.

[10] Christoph Schmidt, Michael Stadtschnitzer, and Joachim Köhler. The Fraunhofer IAIS Audio Mining System: Current State and Future Directions. In *Proceedings of ITG Fachtagung*, Paderborn, Germany, 2016.

[11] Michael Stadtschnitzer, Jochen Schwenninger, Daniel Stein, and Joachim Köhler. Exploiting the large-scale German Broadcast Corpus to boost the Fraunhofer IAIS Speech Recognition System, May 2014.