ASR for Non-standardised Languages with Dialectal Variation: the case of Swiss German

Iuliia Nigmatulina[1], **Tannon Kew**[2], **Tanja Samardžić**[1]
[1]URPP Language and Space, University of Zurich
[2]Department of Computational Linguistics, University of Zurich
{iuliia.nigmatulina, tanja.samardzic}@uzh.ch, kew@cl.uzh.ch

Abstract

Strong regional variation, together with the lack of standard orthography, makes Swiss German automatic speech recognition (ASR) particularly difficult in a multi-dialectal setting. This paper focuses on one of the many challenges, namely, the choice of the output text to represent non-standardised Swiss German. We investigate two potential options: a) *dialectal* writing – approximate phonemic transcriptions that provide close correspondence between grapheme labels and the acoustic signal but are highly inconsistent and b) *normalised* writing – transcriptions resembling standard German that are relatively consistent but distant from the acoustic signal. To find out which writing facilitates Swiss German ASR, we build several systems using the Kaldi toolkit and a dataset covering 14 regional varieties. A formal comparison shows that the system trained on the normalised transcriptions achieves better results in word error rate (WER) (29.39%) but underperforms at the character level, suggesting dialectal transcriptions offer a viable solution for downstream applications where dialectal differences are important. To better assess word-level performance for dialectal transcriptions, we use a flexible WER measure (FlexWER). When evaluated with this metric, the system trained on dialectal transcriptions outperforms that trained on the normalised writing. Besides establishing a benchmark for Swiss German multi-dialectal ASR, our findings can be helpful in designing ASR systems for other languages without standard orthography.

Index Terms: speech recognition, human-computer interaction, Swiss German dialects

1 Introduction

Over the last few years, advancements in speech technology (word error rates (WER) less than 5%, (Chiu et al., 2018; Xiong et al., 2018; Wang et al., 2019)) have lead to an increased demand for automatic speech recognition (ASR) systems outside of a small set of standardised, high resource languages. However, developing ASR systems for non-standard languages and dialects is difficult since, in addition to the lack of a writing standard, one has to deal with a high degree of regional variation coupled with limited training resources. Thus, in such settings, WERs of more than 40% are common (Ali et al., 2017).

Swiss German is a typical case of a non-standard language for which there is a growing interest in ASR technology. The term *Swiss German* describes a family of mutually intelligible Allemanic dialects spoken in the northern two-thirds of Switzerland. Despite carrying the official status of a dialect, Swiss German is spoken in all spheres of oral communication and has long enjoyed a high degree of cultural value and prestige, unlike many other dialects (Hogg et al., 1984). In the past, the linguistic situation in the region has been described as a case of 'medial diglossia' (Kolde, 1981), in which standard German is used for written communication and spoken in formal settings, while Swiss German is primarily used for everyday spoken communication. Being the primary language for approximately five million people in the region, Swiss German has a growing demand for automatic processing and in particular ASR. However, without a single orthography it is difficult to determine not only how to convert speech to text, but also *which* text to convert it to.

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 15–24
Barcelona, Spain (Online), December 13, 2020

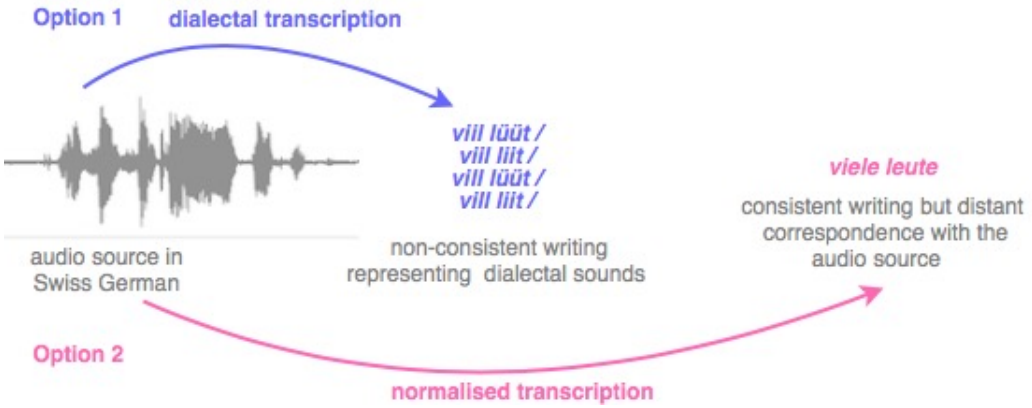

Figure 1: Two potential approaches to Swiss German ASR for the phrase *'viele Leute'* ('many people').

In this paper, we propose the first multi-dialect speech-to-text (STT) framework for Swiss German dialects[1] and explore the efficacy of two potential target textual representations. The first uses *dialectal transcriptions* (option 1 in Figure 1), a writing system proposed by (Dieth, 1986) intended to accommodate all dialects of Swiss German in a manner which closely reflects the sound of the spoken language, while still being easily readable. This is a loosely phonemic spelling system that draws on familiar spelling norms from standard German, restricting the space of possible writings for a given word. It is, however, often subject to transcribers' interpretation leading to considerable intra- and inter-dialectal inconsistency. The second option is to use *normalised transcriptions* (option 2 in Figure 1), a word-level many-to-one mapping between Swiss variants and a single canonical writing resembling standard German whenever possible. Using a normalised transcription should aid ASR by reducing the effect of lexical variability on the target side. On the other hand, it might be harder to establish alignments between the dialectal sound and the standard writing which does not represent the original pronunciation. There is currently no conclusive evidence as to which of the two options yields better results.

The experiments presented in this paper are intended to identify an optimal approach to dealing with Swiss German and non-standard languages in general. The resulting ASR systems, featuring a single unified acoustic model (AM) trained with the time-delay neural network (TDNN) architecture (Waibel et al., 1989) following Kaldi's[2] WSJ *chain* recipe (Peddinti et al., 2015), are the first multi-dialectal solutions for Swiss German and constitute a baseline for future improvements.

2 Related Work

In addressing the topic of multi-dialectal ASR for non-standard languages, our work draws on previous studies carried out in the context of Arabic dialects (e.g. Ali et al. (2014; 2016)). In particular, Khurana et al. (2016) successfully trained a multi-dialect ASR system with a single acoustic model (AM) using approximately 1,200 hours of dialectal Arabic speech data. Leveraging a combination of TDNNs and (bidirectional) long-short term memory networks (bi-LSTMs), trained on speed and volume perturbed data, they reported a WER of 14.7% on held-out test data sampled from different dialects.

In the context of Swiss German, previous studies on ASR have typically focused on one particular regional variety. For example, Imseng et al. (2012) trained an ASR system on the MediaParl corpus, containing Swiss German speech data from the canton of Valais with text transcriptions in standard German. The authors used a triphone-based AM trained with Gaussian mixture models and Hidden Markoc Models (GMM-HMM) and a bigram language model (LM), reporting WER scores of 68.4%. Following this, more advanced AMs have been applied to the same dataset, namely a three-layer ANN-HMM (Razavi et al., 2014), a three-layer CNN-HMM (Palaz, 2016), and a three-layer sMBR discriminative ANN-HMM optimised towards the state-level error rate (Dubagunta and Doss, 2019), with WERs of

[1] `https://github.com/yunigma/Kaldi-for-ASR-of-Swiss-German` (10.10.20).
[2] `http://kaldi-asr.org/doc/index.html` (13.08.20).

25.5%, 23.5% and 18.7%, respectively. In another study, Garner et al. (2014) trained an ASR system on a corpus of radio news broadcasts in Valais German annotated with approximate dialectal transcriptions, achieving 19.4% WER. Note though that the language model used (LM) was trained on the training and evaluation data together and thus failed to provide an accurate estimation of performance on unseen data. Finally, Stadtschnitzer and Schmidt (2018) adapted a standard German ASR system to better handle Swiss German varieties given a small corpus of weather reports, reporting WERs of 23.8% on a challenge corpus of dialectal and standard German speech developed by Baum et al. (2010).

An important aspect of non-standardised-language speech recognition is its evaluation. Usual metrics such as WER and character error rate (CER) assume that only one word or character is correct for each particular position in a sequence. For languages without a standardised orthography, there can be multiple permissible spelling variants for the same word. Therefore, neither WER nor CER are flexible enough to provide an accurate picture of system performance. Ali et al. (2017) proposed WERd (word error rate for dialects) for evaluating dialectal Arabic ASR. They gathered a large collection of potential spelling variants from social media, using a small context window for a given target word or phrase. Valid variants were then found according to occurrence frequencies and normalised with edit distance scores. Then, when comparing the system hypothesis to the reference transcription, a word was considered correct if deemed a valid spelling variant of the corresponding word in the reference. In line with this approach, we exploit the many-to-one mappings provided by the normalised transcriptions in our corpus to better evaluate performance on the dialectal transcriptions.

3 Data

For our experiments, we used the second release of the ArchiMob corpus of Swiss German (Samardžić et al., 2016; Scherrer et al., 2019). The corpus consists of 43 interview recordings with native speakers in 14 different Swiss German dialects, totaling approximately 70 hours of raw speech data.[3] These interviews have been manually transcribed by five native-speaker annotators, with the audio signals aligned to utterance segments typically of around 4-8 seconds in length.

3.1 Transcription Types

The ArchiMob corpus provides two types of transcriptions. Firstly, the Dieth orthography, which constitutes the dialectal transcriptions, is intended to convey the true sound of spoken Swiss German using spelling conventions from standard German (Dieth, 1986). This spelling 'system' essentially guides the writer towards producing relatively consistent spellings, making it easier for anyone familiar with German to read, while still representing the variety and pronunciations of *all* Swiss German dialects. Thus, this orthography can be considered a loosely phonemic textual representation, albeit with considerable variation. Despite the fact that it has not been adopted for use by native speakers, it is commonly used by trained annotators for the purpose of manually transcribing Swiss German speech material according to their intuition of correct pronunciation.

The second type of transcription is a semi-automatically normalised version of the original Dieth transcriptions, where all surface form spelling variants of the same word are mapped to a single normalised form that more closely resembles standard German. This normalisation layer was attained by applying character-level machine translation trained on a small amount of manually normalised transcriptions from the corpus (see Samardžić et al. (2015) , Scherrer et al. (2019)). While this layer by no means constitutes a translation into standard German, it provides a word-level annotation layer that is intended to facilitate automatic processing. Comparing the distinct word types of these two transcriptions shows a considerable effect, reducing the number of distinct word forms (types) from 73,899 to 31,755. Table 1 provides some example instances where multiple Swiss German surface forms in the dialectal transcriptions are mapped to a single standard German form.

[3]Note, dialects are not equally represented in the ArchiMob dataset. The approximate amount of speech data per regional variety is given as follows: Zurich (20hrs), Aargau (10hrs), Bern (8hrs), Lucerne (8hrs), Basel-Stadt (6.5hrs), Glarus (3.3hrs), Uri (1.6hrs), Schwyz (1.6hrs), Valais (1.6hrs), Nidwalden (1.6hrs), Graubünden (1.6hrs), Basel-Landschaft (1.6hrs), Schaffhausen (1.6hrs), St. Gallen (1.6hrs).

Normalised	Dialectal
abbauen	abbaue, abboue, abbuue
abend	aabe, aabed, aaben, aabet, aabid, aabig, abed, abend, abet, abig, abud, obet, obig, oobig, zabig, äbig, òòbed, òòbig
mitbekommen	mitbecho, mitbechoo, mitbichoo, mitbikho

Table 1: Examples of normalisation on dialectal surface variants in the Dieth orthography.

3.2 Data Preparation

For the purpose of our experiments, we split the corpus into training, development and evaluation sets so that they all have a balanced distribution of dialects and speakers. After preprocessing and removing utterances containing overlapping speech signals and anonymised personal names, the sets were reduced to 67,693, 1,710, 1,486 speech utterances, respectively, totalling slightly more than 60 hours of speech data. The small size of the development and evaluation sets was motivated in order to leverage as much data as possible for AM training.

4 Methods

In this section, we describe our approach to building a multi-dialectal ASR system for Swiss German which follows the classical pipeline approach depicted in Figure 2. In particular, we discuss the steps taken to establish a suitable pronunciation lexicon, our choices in training AMs and LMs and the evaluation set up.

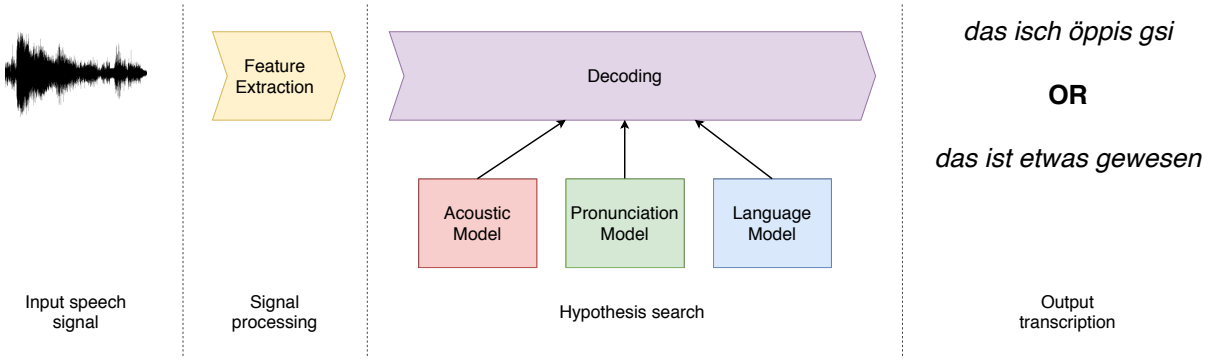

Figure 2: Standard ASR pipeline for Swiss German with potential target textual representations. Note, the output transcription of an input utterance is intended to be either *dialectal* (top) or *normalised* (bottom) depending on the system, not both.

4.1 Pronunciation Lexicon

We took different approaches in establishing pronunciation lexicons for each of the two target textual representations. First, in the case of dialectal writing, we relied on the loosely phonemic representation and derived a pronunciation string for each written word in the training set by segmenting its constituent graphemes with simple heuristics. For example, Swiss German *schwiirigkait* (German *Schwierigkeit*, English 'difficulty') is mapped to the pronunciation string 'sch w i r i g k ai t'.

Second, in the case of normalised writing, we exploited an 11,000-word pronunciation dictionary, mapping standard German words to their Swiss German pronunciations (Schmidt et al., 2020). This dictionary contains manually annotated pronunciation strings (in the SAMPA alphabet (Wells and others, 1997)) for six major regional varieties, namely Zurich, St. Gallen, Bern, Basel, Visp and Nidwalden. To complete the coverage of our data set, we trained a transformer-based a grapheme-to-phoneme (g2p) model[4] on the available pairs (standard German, Swiss SAMPA) and applied it on the words for which

[4] `https://github.com/cmusphinx/g2p-seq2seq` (12.08.20).

manual Swiss SAMPA annotation is missing. We trained the g2p model with the default settings[5] using the pronunciations given for only one of the regional varieties, namely Zurich.

While the overall word-level accuracy of the g2p model, estimated on a held out test set, was only 51.6%, the output still proved useful for the ASR pipeline since it provided good coverage with plausible g2p mappings confirmed by manual inspection of the output. Table 2 shows the percentage of lexical coverage in the train, validation and test sets for each transcription type after these steps.

Split	Dialectal	Normalised
Train	100%	99.9%
Dev	80.2%	82.7%
Test	79.9%	84.8%

Table 2: Pronunciation lexicon coverage on dataset splits. Note, while coverage of the normalised training set is almost complete, a small number of words were missed due to a mismatch in graphemes between the g2p training data and the ArchiMob training set.

4.2 Language Model

The language modeling component in all systems is a statistical N-gram LM. An intrinsic evaluation of several techniques, using test set perplexity (PPL) as a metric, revealed that 3-gram LMs with interpolated modified Kneser-Ney smoothing (Chen and Goodman, 1999), as implemented with MITLM (Hsu and Glass, 2008)[6], consistently yielded the best performance for the dialectal transcriptions. Meanwhile, 5-gram LMs provided additional improvement for the normalised transcriptions. We used these settings for our experiments with varied size of the training data.

The ArchiMob corpus training set provides a total of 76K text utterances which we used for training our base LMs. To test the impact of a larger LM, we created an additional LM trained on the ArchiMob corpus plus some out-of-domain (OOD) data. In order to derive the optimum LM, we gradually concatenated utterances from additional corpora selected for each transcription type individually and trained a series of 3-gram LMs, each time increasing the number of training utterances by 10,000. We first added Swiss corpora and then proceeded with standard German. In order to ensure comparability between LMs, the vocabulary must be consistent (Buck et al., 2014). Therefore, we restricted it to match that of the ArchiMob corpus.

Figure 3 depicts the resulting test set PPL scores with the variable increase in training utterances. As expected, dialectal spelling LMs have considerably higher PPLs than their normalised counterparts. This indicates the challenge of handling a high degree of lexical variety. For both transcription types, however, slight improvements in test set PPL are attained by including a small amount of OOD data from the additional corpora. For dialectal spelling, the minimum PPL is achieved with up to 90,000 utterances, while for normalised text, PPL starts to increase slowly after 80,000 utterances. Following these tests, we selected larger LMs for our ASR experiments, namely the 90k-utterance LM for dialectal writing and the 80k-utterance LM for normalised writing.

4.3 Flexible Evaluation: FlexWER

To score the performance of the systems, besides using the standard metrics of WER and CER, we introduce a soft evaluation measure called FlexWER. The general idea follows Ali et al. (2017), however, we rely on the word-level mapping between the dialectal and the normalised transcriptions provided in the ArchiMob corpus rather than using external sources. Since in our test set we dealt with a closed vocabulary from the corpus, we do not introduce any further metrics, such as an edit distance threshold. Using this measure, a word in the system output hypothesis transcription is considered correct if it shares the same normalised word as the corresponding word in the ground-truth transcription.

[5]Default settings for g2p-seq2seq are as follows: size of each hidden layer = 256, number of layers = 3, size of the filter layer in a convolutional layer = 512, number of heads in multi-attention mechanism = 4.

[6]`https://github.com/mitlm/mitlm` (10.08.20).

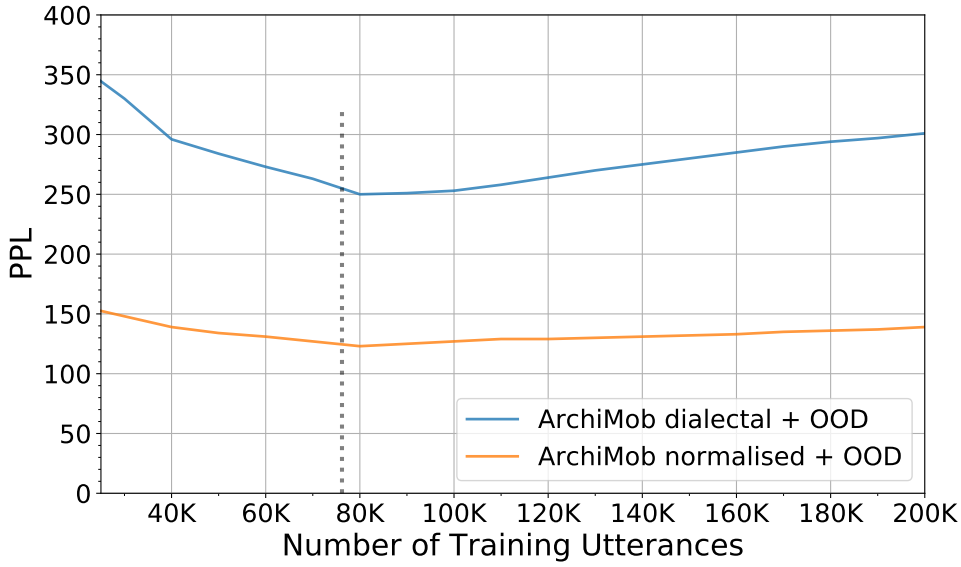

Figure 3: Test set PPL scores for LMs trained on the dialectal transcriptions and normalised transcriptions with additional OOD data. The grey dotted line indicates the upper limit of in-domain utterances (76K).

4.4 Experimental Setup

As a baseline, we built an ASR system based on the WSJ recipe[7] provided in the Kaldi toolkit. We used 13-dimensional Mel-Frequency Cepstral Coefficients (MFCC) features with cepstral mean-variance normalisation (CMVN), the first and second derivatives, and Linear Discriminative Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) transformations. The AM was discriminatively trained using a DNN with state-level minimum bayes risk (sMBR) criterion. The alignment between acoustic signal segments and transcriptions was attained with the GMM-HMM discriminative model with 4,000 senones and 40,000 Gaussians. Since no multi-dialectal ASR baseline exists for Swiss German, we decided to use this simple Kaldi setup as a starting point for our experiments as it allows for comparison between the contributions made by both the AM and LM components. We trained one AM on the dialectal transcriptions and a second AM on the normalised transcriptions and combined these with the pronunciation lexicons and LMs described above to derive our final systems.

For the main models, the TDNN architecture from the WSJ *chain* recipe was adapted (Peddinti et al., 2015) based on the GMM-HMM alignment from Kaldi pipeline. To increase the amount of training data and improve its robustness, we performed audio speed perturbation with speed factors of 0.9, 1.0, 1.1, followed by volume perturbation. In addition to the features used in the baseline model, we also included 100-dimensional iVectors extracted from each speech frame in order to normalise the variation between speakers and dialectal varieties.

In total, we trained six different systems (three for each transcription type): (i) **NNET-DISC-baseLM** (baseline) — discriminatively trained NN AM with the base LM trained on the ArchiMob data only; (ii) **TDNN-iVector-baseLM** — TDNN AM with the base LM; (iii) **TDNN-iVector-dial90k** — TDNN AM with the LM trained on the ArchiMob and OOD data; (iv) **NNET-DISC-baseLM normalised** — similar to the baseline set up but trained with the normalised transcriptions; (v) **TDNN-iVector-baseLM normalised** — similar to the TDNN-iVector-baseLM set up but trained with the normalised transcriptions; (vi) **TDNN-iVector-norm80k** — TDNN AM with the LM trained on the normalised ArchiMob transcriptions and OOD data.

All the systems are evaluated with the standard WER measure on both development and test sets. The systems trained on the dialectal transcription were additionally evaluated with FlexWER measure to account for permissible spelling variation. Finally, we report CER scores too. While this measure

[7] `https://github.com/kaldi-asr/kaldi/tree/master/egs/wsj` (01.05.20).

Transcription	AM	LM	Dev			Test		
			WER	FlexWER	CER	WER	FlexWER	CER
dialectal	NNET-DISC (baseline)	dial-baseLM	54.85	32.87	22.7	54.39	32.09	22.19
dialectal	TDNN-iVector	dial-baseLM	43.18	23.3	15.23	42.38	21.53	14.81
dialectal	TDNN-iVector	dial90k	41.88	21.94	14.97	42.16	**21.27**	**14.64**
normalised	NNET-DISC	norm-baseLM	43.30	–	24.82	40.81	–	23.19
normalised	TDNN-iVector	norm-baseLM	31.96	–	16.97	29.91	–	15.20
normalised	TDNN-iVector	norm80k	31.65	–	16.48	**29.39**	–	14.77

Table 3: WER/FlexWER (where applicable) and CER evaluation of the models.

is typically used when the word boundaries are unclear, we use CER as an indicator of subword-level matches, which are potentially interesting for capturing dialectal variation.

Note that the main goal of our experimental setup is not to establish the advantage of one model over another (e.g. TDNN over NNET-DISC), but to investigate the effect of the writing choice on the target side. To observe this effect, one should compare each dialectal setting with its corresponding normalised setting in Table 3.

5 Results & Discussion

The evaluation results presented in Table 3 indicate that according to the standard WER metric, the best system for multi-dialectal ASR for Swiss German is the one trained on the normalised transcriptions, with a test set WER of 29.39% (see the Appendix for examples of model predictions). Normalised transcriptions effectively reduce the high degree of noise in the transcriptions, allowing for more reliable and robust LMs and also reducing the amount of OOV words in our automatically extended lexicon. The better performance of these models on the test set compared to the development set may be explained by a slightly higher lexicon coverage for the test set: 84.8% vs. 82.7% (see Table 2).

If we consider the FlexWER evaluation measure, which accounts for permissible spelling variants, the best scores are achieved by the dialectal system TDNN-iVector-dial90k (21.27%). The major disadvantage of systems trained on dialectal transcriptions is that the higher degree of lexical variability hinders the contribution of the LM and, for getting more reliable performance estimation, additional normalisation of the results, or even manual evaluation, should be performed. At the same time, the normalised models trained on transcriptions that poorly reflect the pronunciation and thus rely on a simple g2p-based pronunciation lexicon generally miss more subword-level information. According to CER, which is more sensitive to more precise correspondence between the acoustic signal and the annotation than WER, the systems trained on the dialectal transcriptions perform slightly better than their normalised counterparts. This is most noticeable on the development set, where TDNN-iVector-dial90k and TDNN-iVector-norm80k systems score 14.97% and 16.48%, respectively. Therefore, these models may be preferable when dialectal variability is desirable in the output transcriptions.

For the normalised transcriptions, leveraging a slightly larger LM results in only minor improvements in WER over the baseline. We hypothesise that this is a consequence of the restricted domain of the ArchiMob dataset, which limits the effectiveness of any additional information provided by extending LM training data. Further evaluation on OOD data is still needed to answer the question of how well the system is able to adapt to new domains given improved LMs.

The evaluation results can differ between dialects depending on how well a dialect is represented in training data and the degree to which certain dialects differ from the majority of the training data. To attain an impression of these effects, we report WER scores as a distribution over individual dialects (see Figure 4).

Firstly, the effect of the proportion of a corresponding dialect in the training data does not seem to be borne out. The Graubünden dialect has the lowest WER but is represented in the ArchiMob corpus with only a single interview. Similarly, Basel-Landschaft and Schaffhausen varieties, which also have

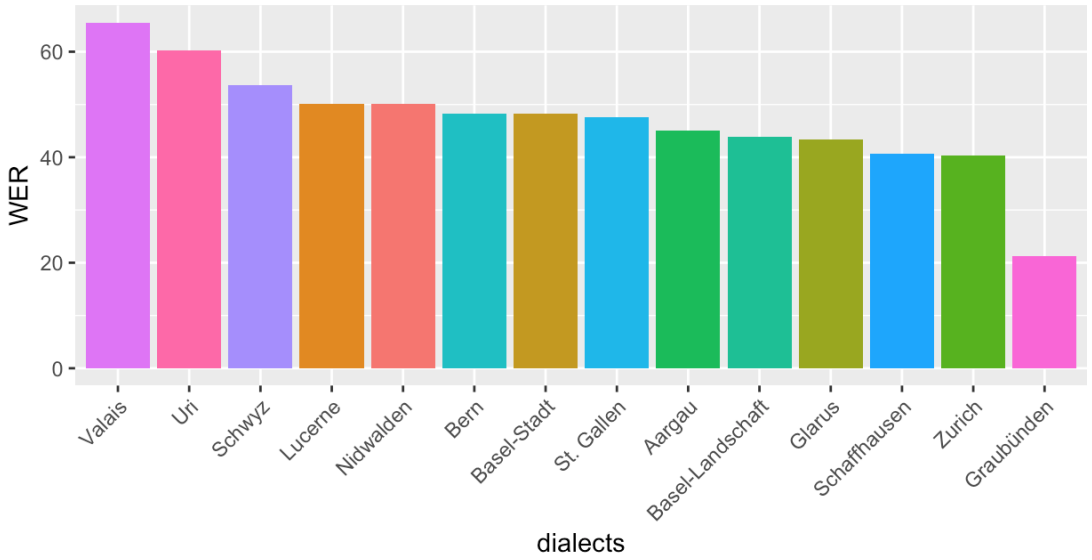

Figure 4: WER of the dialectal TDNN-iVector system evaluated on different dialects.

only one interview each in the corpus, show lower error rates than regional varieties which are well-represented in the corpus (e.g. Aargau, Bern and Lucerne).

Secondly, the degree of similarity between different dialects demands further investigation and is out of the scope of the current paper. Some tendencies, however, can be noticed from the plot: the system performs considerably worse when evaluated on the Valais dialect compared to other dialects. This observation reflects the fact that Swiss German from the canton of Valais is further removed from the other varieties.

Most notably, evaluation on the Graubünden dialect is consistently good, outperforming the results on the Zurich dialect which comprises the largest part of the training data (12 interviews). This peculiarity may be due to the fact that recognition of a dialect, in our case, is very speaker specific. In the ArchiMob corpus, many dialects have only one interview. This means that they are also represented by only a single speaker. Therefore, in order to attain a more reliable picture of dialect-specific speech recognition with a unified ASR system, more speakers are needed for under-represented dialects.

A general limitation of our work is that it uses a relatively closed domain covered by the ArchiMob corpus (personal narratives on a similar topic). To get an idea of how our models perform outside of this domain, we perform an additional OOD evaluation for our dialectal systems using a small data set with appropriate Dieth orthography transcriptions. The results of this evaluation are predictably worse: 59.81% WER and 37.83% FlexWER for the TDNN-iVector-baseLM system. We suspect that potential improvements can be gained by (a) applying advanced training architectures including end-to-end systems; and, of course, (b) increasing the amount of training data to cover more domains and speakers. Since this paper focuses on the efficacy of the target textual representation and the corresponding pronunciation lexicons, we leave further investigations of these factors for future work.

6 Conclusion

In this paper, we have introduced the first multi-dialect ASR framework for Swiss German that exploits a single unified acoustic model. We have tested currently popular techniques for dealing with noisy data and established a solid baseline for future studies in ASR for Swiss German. We have shown a clear benefit of working with normalised transcriptions (provided a suitable pronunciation lexicon is at hand), but also that loosely phonemic dialectal writing yields output potentially interesting for downstream applications specifically targeting regional variation in non-standard languages.

References

Ahmed Ali, Yifan Zhang, Patrick Cardinal, Najim Dahak, Stephan Vogel, and James Glass. 2014. A complete kaldi recipe for building arabic speech recognition systems. In *2014 IEEE spoken language technology workshop (SLT)*, pages 525–529. IEEE.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

Ahmed Ali, Preslav Nakov, Peter Bell, and Steve Renals. 2017. Werd: Using social text spelling variants for evaluating dialectal speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 141–148. IEEE.

Doris Baum, Daniel Schneider, Jochen Schwenninger, Barbara Samlowski, Thomas Winkler, and Joachim Köhler. 2010. Disco-a german evaluation corpus for challenging problems in the broadcast domain. *LREC 2010*.

Christian Buck, Kenneth Heafield, and Bas Van Ooyen. 2014. N-gram Counts and Language Models from the Common Crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland. European Language Resources Association.

Stanley F Chen and Joshua Goodman. 1999. An Empirical Study of Smoothing Techniques for Language Modeling. *Computer Speech & Language*, 13(4):359–394.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE.

Eugen Dieth. 1986. *Schwyzertütschi Dialäktschrift: Dieth-Schreibung*, volume 1. Sauerländer.

S Pavankumar Dubagunta and Mathew Magimai Doss. 2019. Segment-level training of anns based on acoustic confidence measures for hybrid hmm/ann speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6435–6439. IEEE.

Philip N Garner, David Imseng, and Thomas Meyer. 2014. Automatic speech recognition and translation of a swiss german dialect: Walliserdeutsch. In *Proceedings of Interspeech*, number CONF.

Michael A Hogg, Nicholas Joyce, and Dominic Abrams. 1984. Diglossia in switzerland? a social identity analysis of speaker evaluations. *Journal of language and social psychology*, 3(3):185–196.

Bo-June (Paul) Hsu and James R. Glass. 2008. Iterative language model estimation: Efficient data structure & algorithms. In *In Proceedings of the Ninth Annual Conference of the International Speech Communication Association*, pages 841–844, Brisbane, Australia.

David Imseng, Hervé Bourlard, Holger Caesar, Philip N Garner, Gwénolé Lecorvé, and Alexandre Nanchen. 2012. Mediaparl: Bilingual mixed language accented speech database. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 263–268. IEEE.

Sameer Khurana and Ahmed Ali. 2016. Qcri advanced transcription system (qats) for the arabic multi-dialect broadcast media recognition: Mgb-2 challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 292–298. IEEE.

Gottfried Kolde. 1981. *Sprachkontakte in Gemischtsprachigen Städten: Vergleichende Untersuchungen Über Voraussetzungen Und Formen Sprachlicher Interaktion Verschiedensprachiger Jugendlicher in Den Schweitzer Städten Biel/Bienne Und Fribourg/Freiburg*, volume 37. Steiner Franz Verlag.

Dimitri Palaz. 2016. Towards end-to-end speech recognition. Technical report, EPFL.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Marzieh Razavi, Ramya Rasipuram, and Mathew Magimai-Doss. 2014. On modeling context-dependent clustered states: Comparing hmm/gmm, hybrid hmm/ann and kl-hmm approaches. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7659–7663. IEEE.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2015. Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of the 7th Language and Technology Conference*.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob - a corpus of spoken Swiss German. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. Digitising swiss german: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, pages 1–35.

Larissa Schmidt, Lucy Linder, Sandra Djambazovska, Alexandros Lazaridis, Tanja Samardžić, and Claudiu Musat. 2020. A swiss german dictionary: Variation in speech and writing.

Michael Stadtschnitzer and Christoph Schmidt. 2018. Data-driven pronunciation modeling of swiss german dialectal speech for automatic speech recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339.

Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018.

John C Wells et al. 1997. Sampa computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, 4.

Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. 2018. The microsoft 2017 conversational speech recognition system. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5934–5938. IEEE.

Appendix

	Dialectal TDNN-iVector + dial90k
reference	glòüb ìch echli organisiert uf al fäl am aabig **wo de** vatter **hai cho isch** isch aifach **niit me** daa gsii
prediction	glòüb ìch echli organisiert uf al fäl am aabig ******* **oder** vatter ******* ******* **haichoo** isch aifach ******* **nimi** daa gsii
operation	C C C C C C C C C D S C D D S C C D S C C
reference	han ich **den min maa isch drìssgi** gsii und **ìich** äbe ******* **zwaijezwänzgi guet**
prediction	han ich **de mim ja esch driissgi** gsii und **ich** äbe **zwai zwänzgi guet**
operation	C C S S S S S C C S C I S C
	Normalised TDNN-iVector + norm80k
reference	glaube ich ein_klein organisiert auf alle fälle am abend **wo der** vater **heim gekommen ist** ist einfach **nicht** mehr da gewesen
prediction	glaube ich ein_klein organisiert auf alle fälle am abend ******* **oder** vater ******* ******* ******* ist einfach **nichts** mehr da gewesen
operation	C C C C C C C C C D S C D D D C C S C C C
reference	habe ich dann ******* **mein mann** ist dreissig gewesen und ich eben zweiundzwanzig gut
prediction	habe ich dann **meine man ja** ist dreissig gewesen und ich eben zweiundzwanzig gut
operation	C C C I S S C C C C C C C C

Table 4: Examples of predictions on test data by models trained on 1) dialectal and 2) normalised transcriptions. Asterisk symbols are used in cases of model *deletions* and *insertions* instead of absent words.