Modeling Dialectal Variation for Swiss German Automatic Speech Recognition

Abbas Khosravani[1], *Philip N. Garner*[1], *Alexandros Lazaridis*[2]

[1]Idiap Research Institute, Switzerland
[2]Data, Analytics & AI Group — Swisscom AG, Switzerland
name.lastname@{idiap.ch, swisscom.com}

Abstract

We describe a speech recognition system for Swiss German, a dialectal spoken language in German-speaking Switzerland. Swiss German has no standard orthography, with a significant variation in its written form. To alleviate the uncertainty associated with this variability, we automatically generate a lexicon from which multiple written forms of a given word in any dialect can be generated. The lexicon is built from a small (incomplete) handcrafted lexicon designed by linguistic experts and contains forms of common words in various Swiss German dialects. We exploit the powerful speech representation of self-supervised acoustic pre-training (wav2vec) to address the low-resource nature of the spoken dialects. The proposed approach results in an overall relative improvement of 9% word error rate compared to one based on an expert-generated lexicon for our TV Box voice assistant application.

Index Terms: Speech recognition, Wav2vec, dialectal lexicon, Swiss German, multi-dialect, Swisscom, voice assistant, TV Box

1. Introduction

In Switzerland, Standard German (the one spoken across Germany) stands in a diglossic relationship with Swiss German dialects, the varieties of everyday communication throughout the German-speaking cantons. Swiss Standard German (referred to as *Schweizer Schriftdeutsch*) is a variety of Standard German and the written form of the official German spoken in Switzerland. It is used in books, newspapers, and all official publications; however, Swiss Standard German is not typically spoken. Writing in Swiss German has only arisen rather recently (notably in text messaging) and as a result, there are no orthographic conventions for Swiss German varieties.

Swiss Standard German is different from Standard German on all levels of linguistic analysis (called *Helvetisms*) including vocabulary, pronunciation, orthography, and even syntax. Differences at these levels also exist among the various Swiss German dialects. Fortunately, Swiss German is the best researched dialect area in Central Europe [1]. The Dieth spelling system [2] (a system of phonetic transcription) is used in most scientific accounts for writing Swiss German dialects (referred to as GSW). It uses Standard German spelling as a starting point but deviates where it is inconsistent or lacks the precision needed for the description of the various Swiss dialects. Dialectal variation causes lexical units to be pronounced, and therefore written, differently in different regions. For instance, in Standard German *'gehst mir bitte auf das wetter'* ('Tell me the weather please') can be written in Swiss German as *'geisch mer bitte uf ds wätter'* or *'gaischt mer bitte uff daas wetter'* in different regions, where *geisch/gaischt*, *uf/uff*, *ds/daas* and *wätter* are variations of Standard German words *gehst*, *auf*, *das* and *wetter*, respectively. To establish the lexical identity of all writing variants, they need to be normalized to a single form. In this way, whenever a Swiss German word form corresponds to a Standard German form in morphology or etymology (but not the translation), the Standard German form can be used for normalization, otherwise, it is normalized using a common Swiss German form or using the word form that is understood well by the vast majority of the Swiss German speaking people.

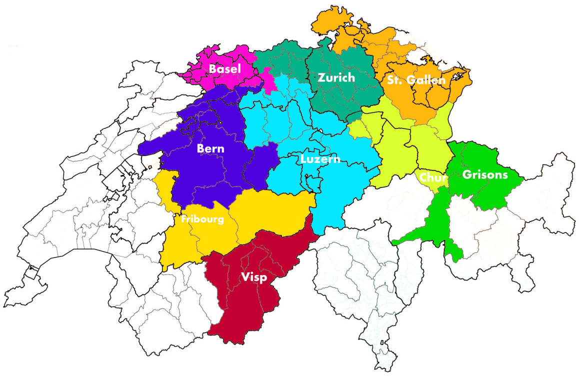

Figure 1: *Swiss German dialectal landscape obtained using cluster analysis of morphological data based on figure 3 of [4].*

We present a multi-dialectal approach to word generation for Swiss German to handle the linguistic variation of words in various dialect landscapes. Given a Swiss German word form and a set of handcrafted written variants, we learn a model to automatically generate inflected word candidates for unseen words or alternative forms for known words in different Swiss German dialects. We utilize the Swisscom Dictionary of spoken and written Swiss German [3], the first publicly described Swiss German dictionary shared for research purposes. It maps 11,248 Standard German words to their spontaneous written form in six Swiss German varieties in Zurich, St. Gallen, Basel, Bern, Visp, and Stans regions. For the majority of the Standard German words, there are manual annotations for Visp and Zurich, but for other regions there is only a small subset available. Figure 1 indicates a geographic classification of dialect areas obtained by cluster analysis of morphological data from linguistic atlas, which is a standard procedure in dialectometry [4].

Automatic speech recognition (ASR) of Swiss German is still a considerable challenge owing to lack of transcribed datasets and the considerable regional variation described above. Exploratory experiments have shown that using a small amount of data in Swiss German varieties is more useful to improve neural network acoustic models than much larger data sets in Standard German [5, 6, 7]. Recently, unsupervised representation learning of speech data by solving a self-supervised context-prediction task [8, 9, 10] has shown to be very effective for speech recognition, particularly on low-resourced languages. We apply a cross-lingual speech representation (XLSR) pre-trained model [8] built on top of wav2vec 2.0 [9], and fine-tune it on labeled multi-dialect speech data in our TV voice assistant domain. We describe Swisscom proprietary multi-dialect speech data designed for TV Box voice assistant and show that by sharing discrete acoustic units across languages we can obtain interesting results on various low-resourced dialects of Swiss German. In the remainder of this paper, we demonstrate two assumptions:

- Transcribing large amounts of data in each dialect is not easy, therefore the situation is akin to that of low-resourced language processing.
- Fine-tuning a pre-trained cross-lingual speech representation model shares capacity across languages and hence should be effective for highly dialectical Swiss German speech data.

We also contribute:

- A method to bootstrap automatic processing using a small handcrafted lexicon and generate Swiss German word candidates in various dialects.
- Improved ASR performance by establishing a lexical identity for writing variants through a dialectal lexicon.

The rest of the paper is organized as follows. Section 2, introduces related work. Section 3 describes the Swisscom lexicon for Swiss German, our approach for dialect variation modeling as well as the ASR system. We describe our proprietary Swiss German speech dataset in TV domain and the ASR results in Section 4. Finally, we conclude in Section 5 with future directions.

2. Related Work

Acoustic modeling for speech recognition typically requires a large volume of training data to properly generalize to a new condition that introduces, e.g., a new accent, recording environment, topic, or emotional state. In real-world scenarios, to reduce the practical demand for supervised data, unsupervised and representation learning techniques have gained a lot of attention recently [11, 12, 13, 14]. Self-supervised representation learning is a new paradigm for end-to-end speech recognition that first learns a general data representation from unlabeled speech data and then fine-tunes on labeled, e.g., low-resourced in-domain, speech data which greatly relieves over-fitting and allows training high capacity models [14, 15, 8]. Cross-lingual speech representation [9], built on top of wav2vec 2.0 [14], has shown to be very effective by sharing discrete acoustic units across languages. Cross-lingual adaptation can be performed by adding a classifier representing character targets plus a word boundary token on top of the model and fine-tuning with a Connectionist Temporal Classification (CTC) loss [16] for speech recognition.

Although extensive research has been performed on Swiss German dialects, some dialect corpora have been collected only recently. *Phonogrammarchiv* of the University of Zurich[1] has a rich but old collection of speech corpora in all Swiss dialects for all national languages. However, the digitization of this archive is an ongoing project [17]. We may find corpora like MediaParl [18], a Swiss German speech dataset from the canton of Valais or SRF Meteo [5], being Swiss German weather reports of SRF Meteo, but they all come with Standard German transcription and have limited domain. The ArchiMob multi-dialectal corpus is the first free general-purpose corpus of spoken Swiss German dialects based on oral history interviews [19, 17]. It initialized the first attempt to provide automatic normalization to deal with the heterogeneous and non-standardized Swiss German spelling [7]. Automatic conversion methods using either character-level statistical machine translation [20, 21, 22], encoder-decoder recurrent networks [23] or recently using transformer networks [3] have also been proposed. In an ASR system, normalization can be performed through a dialectal lexicon which maps dialectal forms of a word to a normalized representation. However, building such a lexicon is normally time consuming and expensive, [3, 19, 24] hence, a grapheme-to-phoneme (G2P) converter [25, 26] is usually trained on a seed handcrafted lexicon to generate multiple pronunciations for new words [5, 6]. Recently, a Swiss German dictionary has been introduced in [3] that maps common words in various Swiss German dialects to Standard German, complemented by phonetic transcriptions in the SAMPA alphabet[2]. However, it is still a limited dictionary and is only partly generated manually. A Transformer-based phoneme-to-grapheme (P2G) model has also been incorporated to generate the most likely spellings based on the pronunciations of the words in each dialect.

It is also desirable to have Standard German as the output of the ASR system [27]. However, dialectal pronunciation is mostly different from the written form of Standard German and translation to Standard German would require substantial syntactic transformations and lexical replacement. To bridge this gap, in [5] a data-driven method was proposed to improve the pronunciation model. The authors show that replacing the missing Swiss German pronunciations with either Standard German phoneme or grapheme sequences is the most effective. They also tried to adapt a Standard German ASR to Swiss German pronunciations by employing a Swiss German G2P model and a phoneme-based language model.

3. Methods

3.1. Lexicon

We use the Swisscom dictionary of spoken and written Swiss German[3], a freely available dictionary for research purposes, which maps Standard German words to Swiss German pronunciations and spontaneous writings [3]. The dictionary includes a total of 11,248 Standard German words with their pronunciation and written representations in six different Swiss dialects: Zurich, Visp, St. Gallen, Basel, Bern, and Stans. To generate pronunciations, we rely on the knowledge of native speakers, dialect-specific grammars and lexica [28], Swiss German dictionaries [1] as well as conferring with friends and acquaintances originating from the respective locations. The non-standard Swiss German writing is partly generated manually by linguists and partly by an automatic process. The manual annotations exist for a subset of 9,000 Standard German words in Zurich and Visp dialects (by native linguists without referring to the pronunciations), and a subset of 600 words for other dialects (by non-native linguists and relying on pronunciations). The automatic process includes training a P2G model based on the Transformer network [29] to map handcrafted pronunciations to their corresponding written form and obtaining the most likely GSW spellings. This results in transcribing 8,000 Standard German words into two Zurich forms and one Visp form.

[1] https://www.phonogrammarchiv.uzh.ch

[2] https://en.wikipedia.org/wiki/X-SAMPA

[3] http://tiny.uzh.ch/11X

Table 1: *Examples of word generation in Swiss German for dialect variation modeling.*

Word	Manual	Automatic
programmieren	programmiere (ZH) programmiärä (ZH) programmieru (VS)	programmierun programmiäre programmierun programmiären
wohnzimmer		wohnzemmer wohzimmer wonzimmer wohnzmmer

3.2. Dialect variation modeling

In this section, we describe our multi-dialectal approach to word generation in Swiss German to handle the linguistic variation of words in various dialect landscapes and hence automatic writing normalization. We only consider variations of lexical units that are considered identical, but are pronounced, and therefore also written, differently in different dialects. As a matter of fact, even experts may not provide consistent annotations of the same text. An end-to-end ASR system which is trained to transcribe in Swiss German dialects may generate character sequences corresponding to a variant of a desired word, either valid or not, which leads to a higher word error rate. Training a language model (LM) on Swiss German texts also results in inconsistency and higher perplexity of the LM. In order to reduce this inconsistency, we need to establish an identity across word variants, e.g., through normalization.

There are multiple normalization strategies available for Swiss German dialects. Normalization may be performed manually [30] or automatically [21, 7], mainly by incorporating statistical machine translation techniques to learn a mapping between original and normalized words. Our approach to normalization is similar to that presented in [17] and is as follows. For all the known words in the handcrafted dictionary, whenever a Swiss German word variant corresponds to a Standard German word in morphology or etymology, the Standard German form is used, otherwise, it is normalized using a common Swiss German form or using the word form that is understood well by the vast majority of the Swiss German speaking people.

After normalization, a grapheme-to-spelling lexicon is generated for training, where every normalized word form is mapped to its character sequence as well as to all its variant forms generated manually. Then we learn to approximate a mapping by first aligning the training lexicon using an Expectation–Maximisation (EM) algorithm and use the alignments to train a standard n-gram model, which is subsequently converted into a Weighted Finite-State Transducer (WFST) model [26]. We utilize Phonetisaurus [26], a G2P conversion toolkit with joint n-gram models in the WFST framework. Since this is a probabilistic model, it is possible to generate multiple written forms of a given word by extracting the *N* best-paths through the graph model. Table 1 provides two examples of the various forms of a word generated using our model. The first word *programmieren* (program) is a known word with available manual GSW forms, but no manual annotation is available for the Standard German word *wohnzimmer* (living room).

The proposed Swiss German word candidates may be a valid dialectal written form or not (in these cases the majority are). However, it does not matter since it is not for a user to know but for the ASR system to normalize word variants when performing a beam-search decoding. Since normalization is context dependent, and the fact that in Swiss German a word could be mapped to multiple normalized forms, a language model plays a crucial rule to reduce ambiguity and hence improve performance in terms of word error rate. We train our LM on both normalized Swiss German text, which represents our target text very well, and Standard German text.

3.3. ASR

We use cross-lingual speech representations (XLSR) [9] built on top of wav2vec 2.0 [8], a framework for self-supervised learning of representations from the raw waveform of speech. The model is publicly available. Cross-lingual representation learning or pre-training aims to build models which leverage unsupervised multilingual data to share discrete acoustic units across languages, particularly for low-resourced languages, creating bridges across languages. It encodes speech audio via a multi-layer convolutional neural network which is then fed to a Transformer network [29] to learn contextualized representations via a contrastive task where the true latent is to be distinguished from distractors [8]. The XLSR model has a large capacity (315M parameters) and contains 24 transformer blocks with model dimensions 1,024, inner dimension 4,096, and 16 attention heads. It is trained on 56k hours of speech data from 53 languages including the CommonVoice (36 languages, 3.6k hours) [31], Babel (17 languages, 1.7k hours) [32], and Multilingual LibriSpeech (8 languages, 50k hours) [33] corpora. We adapt the model by fine-tuning it on our TV domain labeled speech data with a Connectionist Temporal Classification (CTC) loss [16] after adding a classifier representing character targets plus a word boundary token on top of the model. For the first 10k updates only the output classifier is trained, after which the Transformer network is also updated. Models are implemented in Fairseq [34].

We consider a 5-gram word-based language model (LM) trained on the training transcription as well as some augmented text data (either on the original multi-dialect written form or after word normalization) using the KenLM [35] software. To augment text in the TV domain, we manually generate possible text templates when a user interacts with the TV through the box. Then we fill in the templates with various entities, including movie names, city names, etc. We use the beam-search decoder of Wav2letter [36] to generate both lexicon-free as well as lexicon-based transcription for each utterance. The lexicon-free approach utilizes a word separator which is predicted as a normal character during training to split the predicted character sequence into words. In the lexicon-based approach, however, we restrict the beam-search to the words in a specific lexicon and utilize a word-based LM to generate log-probability scores that are then accumulated together with acoustic model scores for a one-pass decoding.

Table 2: *Performance comparison of various multi-dialect ASR systems on the evaluation set.*

			ALL		**GSW Dialects** (WER%)						
System	Lexicon	LM$_{ppl}$	WER%	CER%	NW	BE	ZH	GR	EA	VS	CE
Lexicon-Free	No	-	17.4	4.18	17.2	18.2	16.9	20.8	14.7	19.3	17.4
Baseline	Yes	96.3	16.3	**4.05**	15.8	17.1	15.7	20.4	13.6	17.5	16.3
Manual	Yes	90.8	14.7	4.39	13.7	15.9	13.8	19.6	11.6	16.5	14.8
Automatic	Yes	**83.4**	**13.5**	5.28	**12.6**	**15.1**	**12.2**	**18.1**	**10.4**	**15.7**	**13.7**

4. Experiments

4.1. Data

We describe a proprietary Swiss German multi-dialect dataset designed to improve the Swisscom TV box voice assistant [4]. We aimed at collecting speech data that mimic a realistic usage scenario (e.g. users at different locations in a room talking to the TV box). With the help of 8 different microphones placed at different locations (2 close-talk, 1 pressure zone, 1 mid-field and 4 far-field) we collect user voice commands when interacting with TV assistant [5], e.g. *'what will the weather be like tomorrow in Bern.'* in Swiss German. The recording sessions were either scripted (a speaker reads from a written note) or free talk (speaker could improvise). The transcriptions are generated manually with the help of linguists in various Swiss German dialects. In total, the corpus consists of 392,420 short utterances comprising 440 hours of speech data from 3817 speakers (aged 14 to 89). It covers different Swiss German dialects from different regions including, Bern (BE), Valais (VS), Zurich (ZH), Eastern Swiss (EA), Grisons (GR), Central Swiss (CE) and Northwestern Swiss (NW), however, the distribution of the dialects is somewhat imbalanced. The data is split into train (417h) and validation (23h) sets for system development. We use a different set (23h) for system evaluation.

4.2. Results

In this section, we present the experiments conducted to evaluate the effectiveness of the proposed dialect modeling for Swiss German speech recognition. We start by fine-tuning the XLSR pre-trained model as described in Section 3.3 on the multi-dialect Swiss German training data of the TV application domain. Our experiment indicates that using a cross-lingual pre-trained model significantly outperforms our previous system on low-resourced dialects (not reported here).

Table 2 presents the results. The first system is obtained by lexicon-free decoding without LM. The huge difference between character error rate (CER) and word error rate (WER) indicates the inconsistency in Swiss German word spelling. This system is capable of handling out-of-vocabulary (OOV) words which is important for our application domain. Using a lexicon that maps every word to its spelling without normalization (the baseline system), on the other hand, limits the beam-search decoding to the words in the lexicon. It results in improvement across all dialects with the highest achieved for VS dialect (1.8%) and the lowest for GR dialect (0.4%). For the rest, we observe almost the same improvement. Since we have the lowest amount of speech data for VS (35h) and GR (44h) dialects, it can not be an explanation for this. However, incorporating a word-based LM that is trained on GSW text data (without normalization) and a large variation in spoken Swiss German in the Grisons area can best explain this result (see the results after dialect variation modeling). We should note that only a character-based LM can be used for lexicon-free beam-search decoding since the acoustic model generates character sequences. However, we did not find such an LM beneficial.

The third system (manual) uses a grapheme-to-spelling lexicon that maps every normalized word form to its character sequence as well as to all its manually generated written forms in different dialects. The last system (automatic), on the other hand, incorporates our proposed dialect variation modeling to automatically generate a dialectal lexicon as explained in Section 3.2. Compared to the baseline, the proposed approach provides much better LM perplexity (interpret it as the weighted branching factor) on the evaluation set (83.4 vs. 90.8), resulting in reduced inconsistency of word variation and as a result, improved WER. It provides superior ASR performance across all dialects compared to the baseline system, with the highest improvement for ZH (3.5%), NW, EA (3.2%), and the lowest for VS (1.8%). We also observe a very high improvement of 1.5% for the Grisons region compared to the manual system which indicates the success of dialect variation modeling for dialects with few available handcrafted words. We also observe the lowest improvement for regions of Bern and Valais (0.8%) over the manual system, possibly due to lower variability of spellings and rich handcrafted words in the lexicon by native speakers. Overall, our approach results in a relative improvement of 9% WER compared to the system with a manually generated lexicon.

5. Conclusion & Future Work

We presented a study for automatic processing of the variation associated with Swiss German orthography and applied it to the ASR task. We generated multiple word candidates in different Swiss German dialects, particularly for unseen words, which assisted the ASR system to normalize word variants when performing beam-search decoding. Fine-tuning a pre-trained cross-lingual speech representation model also shares capacity across languages and hence was very effective for low-resourced and highly dialectical Swiss German. Although we were able to overcome the difficulty posed by the variations in written Swiss German, it limits the ability of the ASR system to recognize out-of-vocabulary (OOV) words that happen in real applications. For future work, we aim at solving the OOV issue through a subword-based lexicon that can handle Swiss German written variation at the same time.

[4] https://www.swisscom.ch/en/residential/help/device/blue-tv/voice-assistant.html

[5] https://www.swisscom.ch/en/residential/plans-rates/inone-home/swisscom-blue-tv/functions.html

6. References

[1] H. Christen, E. Glaser, and M. Friedli, *Kleiner Sprachatlas der deutschen Schweiz.* Huber, 2013.

[2] E. Dieth and C. Schmid-Cadalbert, *Schwyzertütschi Dialäktschrift: Dieth-Schreibung.* Sauerländer, 1986, vol. 1.

[3] L. Schmidt, L. Linder, S. Djambazovska, A. Lazaridis, T. Samardzic, and C. Musat, "A swiss german dictionary: Variation in speech and writing," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 2720–2725.

[4] Y. Scherrer and P. Stoeckle, "A quantitative approach to swiss german–dialectometric analyses and comparisons of linguistic levels," *Dialectologia et Geolinguistica*, vol. 24, no. 1, pp. 92–125, 2016.

[5] M. Stadtschnitzer and C. Schmidt, "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)*, 2018.

[6] I. Nigmatulina, T. Kew, and T. Samardzic, "Asr for non-standardised languages with dialectal variation: the case of swiss german," in *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 2020, pp. 15–24.

[7] T. Samardzic, Y. Scherrer, and E. Glaser, "Normalising orthographic and dialectal variants for the automatic processing of swiss german," in *Proceedings of the 7th Language and Technology Conference*, 2015.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[9] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[10] A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," *arXiv preprint arXiv:1911.03912*, 2019.

[11] J. Glass, "Towards unsupervised speech processing," in *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA).* IEEE, 2012, pp. 1–4.

[12] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end asr: from supervised to semi-supervised learning with modern architectures," *arXiv preprint arXiv:1911.08460*, 2019.

[13] Y.-A. Chung and J. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," *Proc. Interspeech 2018*, pp. 811–815, 2018.

[14] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Proc. Interspeech 2019*, pp. 3465–3469, 2019.

[15] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *Proc. Interspeech 2019*, pp. 814–818, 2019.

[16] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[17] Y. Scherrer, T. Samardžić, and E. Glaser, "Digitising swiss german: how to process and study a polycentric spoken language," *Language Resources and Evaluation*, vol. 53, no. 4, pp. 735–769, 2019.

[18] P. N. Garner, D. Imseng, and T. Meyer, "Automatic speech recognition and translation of a swiss german dialect: Walliserdeutsch," in *Proceedings of Interspeech*, no. CONF, 2014.

[19] T. Samardzic, Y. Scherrer, and E. Glaser, "Archimob-a corpus of spoken swiss german," in *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016).* European Language Resources Association (ELRA), 2016.

[20] P.-E. Honnet, A. Popescu-Belis, C. Musat, and M. Baeriswyl, "Machine translation of low-resource spoken dialects: Strategies for normalizing swiss german," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[21] Y. Scherrer and N. Ljubešić, "Automatic normalisation of the swiss german archimob corpus using character-level machine translation," in *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, 2016.

[22] J. Tiedemann, "Character-based psmt for closely related languages," in *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, 2009.

[23] M. Lusetti, T. Ruzsics, A. Göhring, T. Samardzic, and E. Stark, "Encoder-decoder methods for text normalization," in *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, 2018, pp. 18–28.

[24] M.-J. Kolly, A. Leemann, V. Dellwo, J.-P. Goldman, I. Hove, and I. Almajai, "Voice äpp. a smartphone application for crowdsourcing swiss german dialect data," in *The Conference of Digital Humanities 2014*, 2014, pp. 231–233.

[25] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech communication*, vol. 50, no. 5, pp. 434–451, 2008.

[26] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the wfst framework," *Natural Language Engineering*, vol. 22, no. 6, p. 907, 2016.

[27] M. Büchi, M. A. Ulasik, M. Hürlimann, F. Benites, P. von Däniken, and M. Cieliebak, "Zhaw-init at germeval 2020 task 4: Low-resource speech-to-text," in *5th SwissText & 16th KONVENS Joint Conference, Zurich (online), 24-25 June 2020.* CEUR Workshop Proceedings, 2020.

[28] J. Fleischer and S. Schmid, "Zurich german," *Journal of the International Phonetic Association*, vol. 36, no. 2, pp. 243–253, 2006.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[30] E. Stark, S. Ueberwasser, and B. Ruef, "Swiss sms corpus. university of zurich," 2009. [Online]. Available: https://sms.linguistik.uzh.ch/

[31] R. Ardila, 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," *arXiv preprint arXiv:1912.06670*, 2019.

[32] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014).* International Speech Communication Association (ISCA), 2014, pp. 16–23.

[33] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *Proc. Interspeech 2020*, pp. 2757–2761, 2020.

[34] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations).* Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 48–53.

[35] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation.* Association for Computational Linguistics, 2011, pp. 187–197.

[36] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, "Wav2letter++: A fast open-source speech recognition system," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 6460–6464.