Regional Variation in the Performance of ASR Models on Croatian and Serbian

Tanja Samardžić
IDSIA USI-SUPSI,
Lugano, Switzerland
tanja.samardzic@supsi.ch

Peter Rupnik and Nikola Ljubešić
Department of Knowledge Technologies,
Jožef Stefan Institute, Ljubljana, Slovenia
{peter.rupnik|nikola.ljubesic}@ijs.si

Abstract

Regional variation was a limiting factor for automatic speech recognition (ASR) before large language models. With the new technology, speech processing becomes more general, which opens the question of how to use data in similar languages such as Croatian and Serbian. In this paper, we analyse model performance in the currently available train-test scenarios with the goal of better understanding the mutual interference of these two languages. Our findings suggest that better performing models are not very sensitive to the regional variation. Models trained from scratch in one of the languages can give good results on both of them, while fine-tuning large pre-trained multilingual models on smaller data sets does not give the expected results.

1 Introduction

For a long time, automatic speech recognition (ASR) was only possible in very limited domains targetting a particular speaker and a particular topic. This lack of generalisation was famously captured in the comedy sketch *Voice Activated Elevator*[1] by the BBC Scotland Burnistoun show first aired in 2009. The sketch shows two persons trying to activate a lift by pronouncing the number 11. Since the system does not react, they guess that the problem is their Scottish accent and try to mimic other accents ("American", "English") with no success. The sketch captures not only the limitations of the ASR technology at that time but also important societal implications of these limitations.

Regional variation in the ASR performance is certainly present in all languages, but it has not been extensively studied as this kind of research is particularly demanding in terms of resources (both data and computing) and study design (representative sampling), while it is hard to isolate factors that determine the performance.

[1] https://www.bbc.co.uk/programmes/p00hbfjw

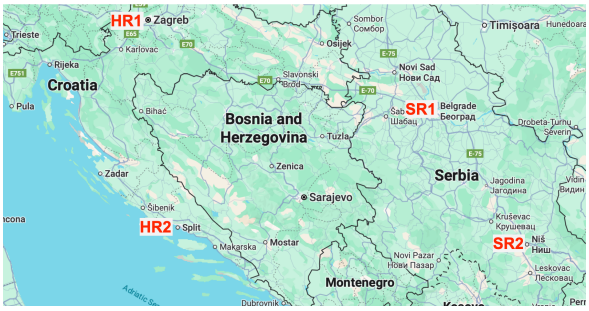

Figure 1: Geographical locations of the four *Mak na konac* data parts. **HR1**: Croatian, Zagreb, Northern variety spoken in the administrative centre. **HR2**: Croatian, Split, Southern variety spoken in a major city. **SR1**: Serbian, Belgrade, Northern variety spoken in the administrative centre. **SR2**: Serbian, Niš, Southern variety spoken in a major city.

New ASR technology that employs large models trained on large amounts of data (thousands of hours of transcribed audio) seems to have reached the level of generalisation that makes multi-purpose ASR viable. In particular, the multilingual model Whisper (Radford et al., 2023) has become a universal reference, able to process various languages, including Croatian and Serbian, without any additional training. Since this model is released as open-weights, it can also be fine-tuned and thus, in theory, more precisely adapted for a particular language and variety.

The possibility of fine-tuning an already well-performing model to reduce its error rate on a given variety is clearly tempting, but the success is not guaranteed. The main problem is a mismatch between the size of the model and the size of the data available for any single variety, this even if we leave aside the problem of computing resources needed to employ a large model.

The goal of our study is to start an exploration towards a better understanding of what happens in various train-test scenarios involving similar but distinct languages and their variants. We focus

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 242–249
March 29, 2026.

on the case of Croatian and Serbian as two languages that both belong to the same macro language, BCMS (for Bosnian, Croatian, Montenegrin and Serbian). The crucial question for macrolanguages is whether the data in one variety can be used to train or improve ASR models in all varieties. Through several recent projects, Croatian and Serbian have been equipped with the necessary resources to start investigating this problem, making them an interesting case study potentially relevant to many other similar languages.

Data part	Source	Language	Region	Size
HR1	Radio Student	Croatian	Zagreb	5h
HR2	TV Dalmacija	Croatian	Split	5h
SR1	Peščanik	Serbian	Belgrade	5h
SR2	Južne vesti	Serbian	Niš	5h

Table 1: Summary of the data key features.

2 Related Work

Regional variation is, to some degree, addressed in in the case of Arabic and Swiss German, which are both famous for their regional diversity.

While Modern Standard Arabic (MSA) has been the main focus of research in the case of Arabic (Dhouib et al., 2022), dialect variation is starting to attract more attention in recent studies (Mubarak et al., 2021; Alharbi et al., 2024; Al-Fetyani et al., 2021; Al Ali and Aldarmaki, 2024; Djanibekov et al., 2025). What emerges from these studies is considerable variation in the ASR performance across dialects. While dialect-specific fine-tuning can help, this applies only if there is enough data (over 800 hours of transcribed speech). Also the impact of data availability does not seem to be straightforward. For instance, the performances tend to be better on the Syrian variety than on the Saudi or Egyptian one although the latter two are far better represented in the available data and models (Djanibekov et al., 2025).

Interesting dialectal patterns can be seen in the studies on Swiss German too: ASR performances are the worst for the Wallis variant while they are the best for the Grisons variant. This same outcome was achieved by two independent studies that evaluated very different ASR systems trained on different data sets. The first evaluation (Nigmatulina et al., 2020) was performed in 2020 with a Swiss-German Kaldi (Povey et al., 2011) recipe trained on the ArhciMob corpus (Samardžić et al., 2016; Scherrer et al., 2019). The second evaluation was performed in 2022 (Schraner et al., 2022) with a multilingually pre-trained XLS-R model (Babu et al., 2021) fine-tuned on other Swiss German data, including SDS-200 (Plüss et al., 2022) and a Swiss local parliaments corpus (Plüss et al., 2021), and tested on the STT4SG-350 data set (Plüss et al., 2023). Both of these variants are geographically peripheral, but one of them seems to be easier for ASR models than the other.

Until recently, the performance of speech-to-text models on Croatian and Serbian was rarely reported. A Kaldi recipe for Serbian (Popović et al., 2015) showed a large performance drop on the out-of-domain test set. In the past few years, however, considerable progress has been made in data development (Ljubešić et al., 2024a,c,b; Rupnik and Ljubešić, 2022a), which enabled model training and testing initially on Croatian (Ljubešić et al., 2022) and more recently on Serbian (Sagić, 2023). These data sets contain over 3000 hours of automatically aligned transcribed audio for Croatian and almost 1000 hours for Serbian.

The project *Mak na konac* (Samardžić et al., 2024) resulted in a new test set consisting of 20 hours of speech sampled from four media sources and manually transcribed. This data set targets specifically regional variation which makes it especially suitable for our analysis.

3 Data

The *Mak na konac* data set (MNK) (Samardžić et al., 2024) is composed of four parts of equal size (5 hours of audio each), two representing Croatian (HR) and two Serbian (SR) varieties. As shown in Figure 1, each data part represents one region. In addition to the administrative centres (Zagreb and Belgrade), both language varieties are represented with a Southern variety (Split and Niš). All data sources listed in Table 1 are interview-type TV or radio programmes featuring one host and one guest speaker, with the exception of Radio Student hosted by two presenters.

This is a multi-reference data set, where each segment of speech is aligned with multiple text reference options created to cover known variability

Model	Pre-training	Training / Fine-tuning
WhisperV	Multilingual	None
WhisperS	WhisperV	Serbian, around 70h: Flores ASR (multilingual Eastern Europe without Croatian) + Serbian Common Voice + JuzneVesti-SR
WhisperSJV	WhisperV	Serbian, around 120h: Flores ASR (multilingual Eastern Europe without Croatian) + Serbian Common Voice + Serbian "Juzne vesti" Serbian Common Voice + Unknown Serbian
Transducer	None	Croatian, 1816 hours, ParlaSpeech.v1
CTC	None	Croatian, 1816 hours, ParlaSpeech.v1
W2V2XLSR	Multilingual only audio: BABEL, Multilingual LibriSpeech, CommonVoice (no Croatian), VoxPopuli (no Serbian)	Croatian, 300 hours, ParlaSpeech.v1
W2V2Slavic	Multilingual only audio: VoxPopuli (no Serbian)	Croatian, 300 hours, ParlaSpeech.v1

Table 2: Summary of key features of the models tested in the MNK project.

in speech transcription. In some options, elements of speech (repetition, fillers) are included in the reference transcript, while they are excluded in others. Further options are created by a different treatment of numbers and abbreviations. As an example, Table 4 shows how such a reference would look like in English. Taking the best score out of all options when evaluating models has the effect of neutralising irrelevant variation and enables comparing models trained on different data. This is especially important when comparing solutions based on off-the-shelf models because the user has no control over the selection of pre-training data.

The audio part of the MNK test data is available online on Hugging Face.[2] The multi-reference transcript is not publicly available, to prevent model contamination, but researchers can obtain an evaluation report on a submitted output of a model.

4 Models

We analyse the performance of seven freely available models:

- WhisperV: multilingual Whisper large-v3, which is pre-trained on 1 million hours of weakly labelled data and 4 million hours of pseudo-labelled data, produced with its predecessor, Whisper-large-v2. It is capable of automatically determining the language of the input speech as well as translating input speech into a variety of languages. In this setting, the model is applied on the *Mak na konac* data in a **zero-shot** fashion.

- WhisperS and WhisperSJV: two variants of the WhisperV model fine-tuned on transcribed Serbian audio. The first variant is fine-tuned on Mozilla Common Voice 13 and Google Fleurs and the ASR training data set for Serbian JuzneVesti-SR v1.0 (Rupnik and Ljubešić, 2022b). The second variant has 50+ hours of unspecified Serbian data added to the training set. Both of these models can be considered fine-tuned towards **Serbian** since Common Voice did not contain Croatian at the time of fine-tuning.

- Transducer and CTC are parts of NVIDIA's NeMo toolkit. The main difference between the two variants is that Transducer takes previously generated letter as input at the next step, while CTC does not (it combines the acoustic and the language model in a more traditional way). In both of these settings, the models that were trained from scratch on **Croatian** parliamentary data set ParlaSpeech-HR (Ljubešić et al., 2022).

- W2V2XLSR and W2V2Slavic are models based on the wav2vec2 architecture. This means that they are first trained on large quantities of unlabelled audio alone then continued training on labelled data. The former is pre-trained on multilingual partially labelled text and speech datasets (XLS-R), while for the latter, the VoxPopuli Slavic labelled data (Wang et al., 2021) are used. For our study, it is important to note that the VoxPopuli data set is part of both settings and that it contains only Croatian without Serbian. Both of these models are finally fine-tuned on a subset of the **Croatian** ParlaSpeech-HR data set.

[2] https://huggingface.co/datasets/classla/mak_na_konac/viewer/default/SR1?row=4

Segment ID		Whisper V	Whisper S	Whisper SJV	Transducer	CTC	W2V2 XLSR	W2V2 Slavic
170918_13_14	r1	[0.286,	[0.429,	[0.429,	[0.571,	[0.571,	[1.000,	[0.857,
	r2	0.143,	**0.286**,	0.286,	0.429,	0.429,	0.857,	0.857,
	r3	0.125,	0.500,	0.250,	0.375,	0.375,	0.750,	0.714,
	r4	**0.000**]	0.375]	**0.125**]	**0.250**]	**0.250**]	**0.625**]	**0.500**]
170918_14_15	r1	[0.429,	[**0.286**,	[0.714,	[0.857,	[0.857,	[0.857,	[0.857,
	r2	**0.400**]	0.400]	**0.300**]	**0.300**]	**0.300**]	**0.300**]	**0.300**]
170918_15_16	r1	[0.111,	[**0.000**,	[0.556,	[0.667,	[0.556,	[0.556,	[0.556,
	r2	**0.091**]	0.182]	**0.455**]	**0.364**]	**0.455**]	**0.364**]	**0.455**]
170918_16_17	r1	[**0.100**,	[**0.100**,	[**0.100**,	[**0.300**,	[**0.400**,	[**0.100**,	[**0.300**,
	r2	**0.100**]	**0.100**]	**0.100**]	**0.300**]	**0.400**]	**0.100**]	**0.300**]
170918_17_18	r1	[0.222,	[0.111,	[0.222,	[0.556,	[0.667,	[0.444,	[0.444,
	r2	0.455,	0.364,	0.455,	0.455,	0.545,	0.636,	0.545,
	r3	0.111,	**0.000**,	0.111,	0.444,	0.556,	0.333,	0.333,
	r4	0.364,	0.273,	0.364,	0.364,	0.455,	0.545,	0.455,
	r5	0.100,	0.200,	0.100,	0.400,	0.500,	0.300,	0.300,
	r6	0.333,	0.417,	0.333,	0.333,	0.417,	0.500,	0.417,
	r7	**0.000**,	0.100,	**0.000**,	0.300,	0.400,	**0.200**,	**0.200**,
	r8	0.250]	0.333]	0.250]	**0.250**]	**0.333**]	0.417]	0.333]

Table 3: The first five entries in an evaluation report. Each entry contains multiple WER scores (one for each reference option r1–rn) for one audio segment. The best scores are **in bold**. The full evaluation report for all data parts contains around 11'000 of such entries.

r1	We're gonna meet on the 20th of December.
r2	We're gonna meet on the twentieth of December.
r3	We are going to meet on the 20th twentieth of December.
r4	We are going to meet on the twentieth of December.

Table 4: An English illustration of a multi-reference data set. Some versions (r1, r2) are more literal and some (r3, r4) more standard. In some versions (r1, r3) the number is written as a digit, while in others (r2, r4) they are spelled out.

These models can be divided into three groups depending on whether they are pre-trained and used in zero-shot testing (WhisperV), pre-trained and fine-tuned for a specific variant (WhisperS, WhisperSJV, W2V2Slavic and W2V2XLSR to some degree) or trained from scratch (Transducer and CTC). The latter option allows more control over the training data and fitting the models more closely to the target, but these models are smaller and might not capture all the nuances that can be accommodated by the pre-trained models that are bigger. On the other hand, the bigger models require much more data to be trained, which is why they cannot be trained for any single language (except English) specifically.

A few notes are due on fine-tuning. The only model that is not specifically fine-tuned for any Croatian or Serbian target data is WhisperV, although it is obvious that some Croatian and Serbian data are included in its, highly multilingual, training set given mid-tier performance on these languages. Two models (WhisperS and WhisperSJV) are said to be fine-tuned towards Serbian, but some of the fine-tuning data might have already been included in the pre-trained model as well. The other models are trained on Croatian data (recordings of the Parliament sessions). Transducer and CTC are trained from scratch on Croatian data, while W2V2XLSR and W2V2Slavic are fine-tuned on a portion of the the same Croatian data. The latter two models are pre-trained on multilingual and Slavic data respectively. A summary of the models' key features is given in Table 2

The models are initially tested in the *Mak na konac* project (Samardžić et al., 2024), which has also provided the multi-reference gold standard. The full evaluation reports created in the original evaluation are large tables that contain meta-data and CER (character error rate) scores in addition to the WER (word error rate) scores shown in Table 3. For the current study, we use only some parts of these reports, the WER scores. We extract the best WER score for each model for a given audio segment and calculate the average best score for each model and each data part. This allows us to observe regional variation in best WER scores per

model.

Our analysis is performed on the evaluation reports for the seven models listed in Table 2 separately for the four data parts described in Table 1. The first five lines of one of the analysed evaluation reports is shown in Table 3. For each audio segment (ID in the first column) and each tested model, we obtain an array of WER scores (here rounded up to three decimal places). For example, the model WhisperV obtains eight scores on the segment 170918_17_18 depending on the overlap between the model output and each reference option. The number of scores per audio segment and per model can go over 70 depending on the presence of numbers and abbreviations, but also on the length of the utterance.

5 Results and Discussion

Figure 2 shows the average best WER score for each model and each data part.

5.1 Zero-Shot Whisper Wins, for Now

It turns out that, somewhat surprisingly, the best performing model overall is WhisperV, followed by the two models trained on Croatian (Transducer and CTC), while the worst performance is obtained with the two W2V2 models fine-tuned on Croatian. It seems that the currently available data for both Croatian and Serbian do not reach the level that is needed to see the benefit of training and fine-tuning specific to the target variety, as it could be seen in Arabic. This might change very soon but, for now, zero-shot use of pre-trained Whisper gives the best results on both Croatian and Serbian.

5.2 Regional Patterns

The only settings that show regional variation are the two versions of Whisper fine-tuned on Serbian (WhisperS and WhisperSJV). In these cases, fine-tuning does not improve the results for Serbian compared to the zero-shot setting, but it does spoil the performance on Croatian. We also note that one part of the data used for fine-tuning comes from the same source as the SR2 data part. Although there is no data overlap, one would expect to see more benefits of fine-tuning on the same data source. In reality, the difference in favour of the zero-shot setting increases on SR2 compared to SR1, which remains puzzling. It looks like fine-tuning resulted in overfitting to the specific data set used for fine-tuning, so the model became more sensitive to any variation.[3] Nevertheless, the drop in the performance is bigger on Croatian.

5.3 Training From Scratch Better Than Fine-Tuning

The only case where zero-shot Whisper does not give the best performance are the two models trained from scratch on Croatian (Transducer and CTC) and tested on HR2. This said, these models show much more variation across data parts with a considerable performance drop on the other Croatian part (HR1), even bigger than on the two Serbian parts. Still, their performance remains superior to all fine-tuned settings on Croatian and better than the two W2V2 models on both languages. These results suggest that training from scratch is a better way for reaching good performance on a target variant at the expense of some generalisation. Fine-tuning large models on smaller data sets seems to result in strong overfitting preventing improvements even on similar data and reducing the transfer across variants.

5.4 Style Variation Rather Than Regional

Another outcome that is unexpected is overall better performance on Southern variants (HR2 and SR2) than on the ones spoken in the two administrative centres (HR1 and SR1). Given the biases in the population size and overall media presences in favour of the central variants, one would expect that they are better represented in the data available for training and fine-tuning models, which, in turn, should lead to better results. Still the pattern is the opposite. The difference is more pronounced in the case of Croatian (HR1 vs. HR2) than in the case of Serbian (SR1 vs. SR2). The data part that seems the hardest for all the models is HR1, while HR2 seems the easiest for all models except the two Whisper variants fine-tuned on Serbian (WhisperS and WhisperSJV). All models trained and fine-tuned on Croatian data perform worse on HR1 than on the two Serbian parts.

This naturally raises the question of what makes HR1 harder than HR2 for all the models regardless of the training/fine-tuning settings. The fact that the source of the data is a radio programme with two hosts and a rather informal conversational setting might be a part of the answer as this might introduce more dynamics in the interactions and more

[3] Note also that the performance reported on the Hugging Face repository (Sagić, 2023) is much better than what we observe, which might also be interpreted as a sign of overfitting.

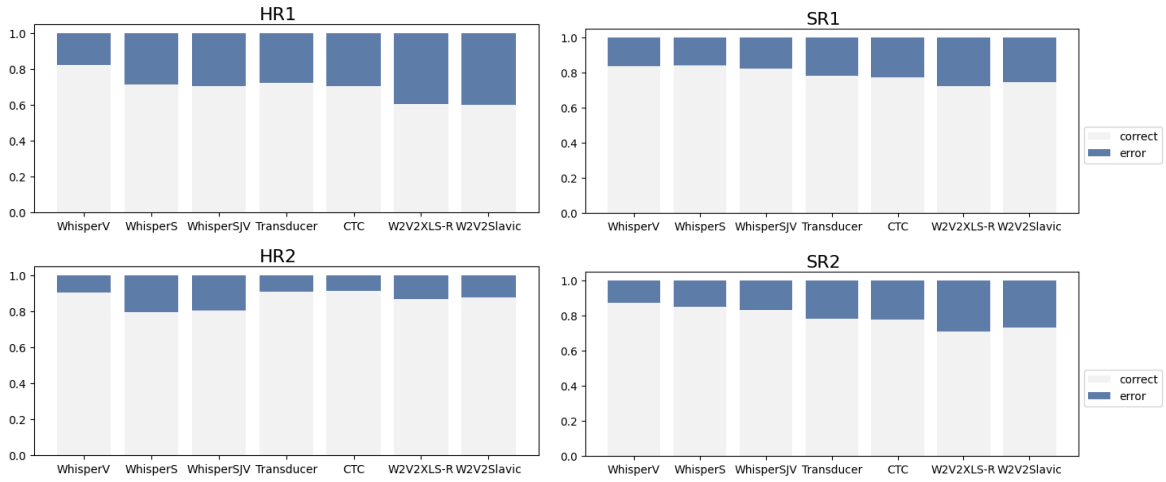

Figure 2: Average best WER scores on the four parts of the MNK data set. The dark bars on the top (error) show the WER, the light bars in the bottom show the complement (1-WER), which can be seen as word accuracy.

small speaker overlaps. Modern ASR models are rather robust to such overlaps, but they still might be impacted. Other explanations might be obtained from the metadata (e.g. the speakers' demographics) and data analysis (e.g. lexical diversity), which remain outside of the scope of our current analysis.

One feature that can be extracted from our data is the level of variability of the WER scores. If we count the number of distinct WER scores assigned to the models, this can show how variable reference transcripts were in each data part. For example, the total number of scores assigned to WhisperV in Table 3 is 18 and the number of distinct scores is 14. Table 5 shows the counts of distinct scores for all data parts. Indeed, model performance seems correlated with the level of variability in the data: the number is the highest in HR1, followed by SR1, then SR2 and HR2. We see the same ranking in the overall model performance. This means that at least some differences in the model performance can be explained by the presence of elements of speech, numbers and abbreviations in the data sets, which can be associated with styles of speech rather than geographical regions.

Model	HR1	HR2	SR1	SR2
WhisperV	380	250	323	250
WhisperS	406	257	315	249
WhisperSJV	427	281	331	268
Transducer	397	266	331	280
CTC	412	272	362	282
W2V2XLS-R	429	273	349	303
W2V2Slavic	436	264	336	302
Average (rounded)	412	266	335	276

Table 5: The number of distinct WER scores per model and data part.

6 Conclusion and Future Work

In this paper, we have studied different scenarios of training and fine-tuning speech-to-text models on Croatian and Serbian with the goal of understanding whether the two variants need to be separated to achieve better overall results.

Our findings suggest that regional variation does play a major role in model's performance in the currently available data and model settings. Composing a large data set that contains both Croatian and Serbian seems to be the best way to achieve good scores on both languages. With additional Serbian data added to the Croatian set, training from scratch might give better scores than zero-shot use and fine-tuning multilingual pre-trained models.

The performance of fine-tuned models in the settings analysed in our study was clearly worse than the other two options (training from scratch and zero-shot testing). While this is a consistent pattern in our findings, the potential of fine-tuning a large pre-trained model is still to be clarified in future work. For this, we will need to evaluate models trained from scratch and fine-tuned on the same amount of data.

Limitations

To study the impact of regional variation on the performance of ASR models on Croatian and Serbian, we used a balanced data set representing four

variants. However, the model training and fine-tuning settings could not be controlled due to limited resources. As a consequence, our settings are not fully comparable. On one hand, we do not have results of the Whisper model fine-tuned on Croatian to compare it with the one fine-tuned on Serbian. On the other hand, we have no results showing what happens if we train from scratch on Serbian or on a combination of Croatian and Serbian. Finally, W2V2 models are fine-tuned only on Croatian. Observing all the settings in a fully comparable way would make our conclusions more sound, but the partial observations that were possible by reusing existing information already show interesting patterns that can be helpful for designing fully comparable settings in future studies.

Another limitation of our study is leaving aside the question of model response time, which is an important issue in speech processing. Most of the models that we analysed are likely to respond too slowly for a practical application. With the available data, we could not analyse this aspect in detail, but we could show that models trained from scratch can outperform large pre-trained models, which is an interesting point for future research taking into account practical aspects such as response time as well.

Finally, our study is performed on data and models that were available at the time of the original evaluation. In the meantime, new data sets were published increasing significantly the size of available data for both Croatian and Serbian. Also a few new models are emerging as good candidates for studies such as ours. We believe that, despite this limitation, our insights can inform future studies regarding the factors to be tested with new data and models.

Acknowledgments

References

Maryam Khalifa Al Ali and Hanan Aldarmaki. 2024. Mixat: A data set of bilingual emirati-English speech. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 222–226, Torino, Italia. ELRA and ICCL.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2021. Masc: Massive arabic speech corpus.

Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim, Raghad Aloraini, Raneem Alnajim, Ranya Alkahtani, Renad Almuasaad, Sara Alrasheed, Shaykhah Alsubaie, and Yaser Alonaizan. 2024. Sada: Saudi audio dataset for arabic. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286–10290.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296.

Amira Dhouib, Achraf Othman, Oussama El Ghoul, Mohamed Koutheair Khribi, and Aisha Al Sinani. 2022. Arabic automatic speech recognition: A systematic literature review. *Applied Sciences*, 12(17).

Amirbek Djanibekov, Hawau Olamide Toyin, Raghad Alshalan, Abdullah Alatir, and Hanan Aldarmaki. 2025. Dialectal coverage and generalization in Arabic speech recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29490–29502, Vienna, Austria. Association for Computational Linguistics.

Nikola Ljubešić, Danijel Koržinek, and Peter Rupnik. 2024a. Parliamentary spoken corpus of Croatian ParlaSpeech-HR 2.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, and Ivo-Pavao Jazbec. 2022. ParlaSpeech-HR - a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 111–116, Marseille, France. European Language Resources Association.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, Ivo-Pavao Jazbec, Vuk Batanović, Lenka Bajčetić, and Bojan Evkoski. 2022. ASR training dataset for croatian ParlaSpeech-HR v1.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Peter Rupnik, and Danijel Koržinek. 2024b. Parliamentary spoken corpus of Serbian ParlaSpeech-RS 1.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Peter Rupnik, and Tea Perinčić. 2024c. The "Mići Princ" text and speech dataset of Chakavian micro-dialects. Slovenian language resource repository CLARIN.SI.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI aljazeera speech resource a large scale annotated Arabic speech corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2274–2285, Online. Association for Computational Linguistics.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. 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*, pages 15–24, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus. *Preprint*, arXiv:2010.02810.

Branislav Popović, Stevan Ostrogonac, Edvin Pakoci, Nikša Jakovljević, and Vlado Delić. 2015. Deep neural network based continuous speech recognition for serbian using the kaldi toolkit. In *Speech and Computer*, pages 186–192, Cham. Springer International Publishing.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Peter Rupnik and Nikola Ljubešić. 2022a. ASR training dataset for serbian JuzneVesti-SR v1.0. Slovenian language resource repository CLARIN.SI.

Peter Rupnik and Nikola Ljubešić. 2022b. ASR training dataset for Serbian JuzneVesti-SR v1.0. Slovenian language resource repository CLARIN.SI.

Andrija Sagić. 2023. Whisper-large-v3-sr-combined.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob - a corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066, Portorož, Slovenia. European Language Resources Association (ELRA).

Tanja Samardžić, Peter Rupnik, Mirjana Starović, and Nikola Ljubešić. 2024. Mak na konac: A multi-reference speech-to-text benchmark for croatian and serbian. Institute of Contemporary History.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. Digitising Swiss German: how to process and study a polycentric spoken language. *Language Resources & Evaluation*, 53:735–769.

Yanick Schraner, Christian Scheller, Michel Plüss, and Manfred Vogel. 2022. Swiss german speech to text system evaluation. *Preprint*, arXiv:2207.00412.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Miguel Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *CoRR*, abs/2101.00390.