Proceedings of ParlaCLARIN III @ LREC2022 pages 111–116
Marseille, 20 June 2022

ParlaSpeech-HR – a Freely Available ASR Dataset for Croatian Bootstrapped from the ParlaMint Corpus

Nikola Ljubešić[1,2], Danijel Koržinek[3], Peter Rupnik[1], Ivo-Pavao Jazbec
[1]Jožef Stefan Institute, Slovenia
[2] Faculty of Computer and Information Science, University of Ljubljana, Slovenia
[3] Polish-Japanese Academy of Information Technology, Poland
nikola.ljubesic@ijs.si, danijel@pja.edu.pl, peter.rupnik@ijs.si, ipjazbec@gmail.com

Abstract

This paper presents our bootstrapping efforts of producing the first large freely available Croatian automatic speech recognition (ASR) dataset, 1,816 hours in size, obtained from parliamentary transcripts and recordings from the ParlaMint corpus. The bootstrapping approach to the dataset building relies on a commercial ASR system for initial data alignment, and building a multilingual-transformer-based ASR system from the initial data for full data alignment. Experiments on the resulting dataset show that the difference between the spoken content and the parliamentary transcripts is present in ∼4-5% of words, which is also the word error rate of our best-performing ASR system. Interestingly, fine-tuning transformer models on either normalized or original data does not show a difference in performance. Models pre-trained on a subset of raw speech data consisting of Slavic languages only show to perform better than those pre-trained on a wider set of languages. With our public release of data, models and code, we are paving the way forward for the preparation of the multi-modal corpus of Croatian parliamentary proceedings, as well as for the development of similar free datasets, models and corpora for other under-resourced languages.

Keywords: parliamentary data, automatic speech recognition, free language resources, Croatian language

1. Introduction

In recent years we have witnessed huge advances in speech technology, primarily by applying the self-supervision paradigm over raw speech data. (Baevski et al., 2020) The new paradigm allows for good ASR systems to be built only with a few tens of hours of speech segments and corresponding transcripts. (Babu et al., 2021)

For Croatian, or any other closely-related language from the HBS macro-language group, including also Bosnian, Montenegrin and Serbian, there are, sadly, no freely available ASR datasets or systems. There has been work on ASR for the HBS macro-language (Martinčić-Ipšić et al., 2008; Popović et al., 2015; Nouza et al., 2016), but none resulted in an open dataset or system, stalling the development of speech technologies for these languages significantly.[1]

Parliamentary proceedings are a very well known source of speech data with already available transcripts (Mansikkaniemi et al., 2017; Helgadóttir et al., 2017; Kirkedal et al., 2020; Solberg and Ortiz, 2022) due to a significant number of countries making the transcripts and recordings freely available under a public license.

For exploiting resources consisting of speech recordings and their transcripts, the transcripts have to be aligned to the speech recordings, which is often a technical challenge due to the amount of data available and no or only high-level alignment between the two modalities. There exists a well established methodology in performing such alignment by using an existing ASR system to automatically transcribe the speech recordings, to align the automatic transcripts from the ASR system with the previously available human transcripts, obtaining thereby alignments between speech recordings and human transcripts. (Katsamanis et al., 2011; Marasek et al., 2014; Panayotov et al., 2015)

This work describes the bootstrapping approach to building the first freely available ASR dataset for Croatian from the parliamentary proceedings available through the ParlaMint corpus (Erjavec et al., 2022). Given the lack of any datasets or models for Croatian and related languages prior to the start of our efforts, the described approach consists of two phases - aligning a smaller amount of data through a commercial ASR system, and then training an in-house ASR system for the alignment of all available data. With this approach the costs of the construction process are kept at their bare minimum.

The main contributions of this paper are the following: we share the first freely available ASR dataset for Croatian, and a methodology for building ASR datasets for other under-resourced languages from parliamentary data. We present our insights into the quality of parliamentary transcripts that are known to deviate from the speech recordings. We investigate the necessity of performing data normalization prior to training state-of-the-art ASR models from parliamentary data. Finally, we map the path forward in creating a multimodal cor-

[1]During our work on ParlaSpeech, the VoxPopuli dataset (Wang et al., 2021) was released, containing parliamentary debates from the European parliament, including also Croatian with 42 hours of transcribed speech. The transcription has, however, significant issues with one half of the text missing non-ascii characters.

pus of Croatian parliamentary proceedings, and call out the parliamentary data community to join the ParlaSpeech initiative in building ASR datasets and multimodal corpora for other under-resourced languages.

2. Dataset Construction

The dataset presented in this paper was constructed in a bootstrapping manner, first constructing a 72-hours corpus by exploiting the commercial Google Speech-To-Text (STT) system[2], then training an in-house ASR system on that corpus, and finally applying that ASR system over all available recordings of the Croatian parliamentary proceedings from the ParlaMint corpus (Erjavec et al., 2021; Erjavec et al., 2022).

The textual source of the data is the Croatian portion of the ParlaMint corpus, containing the proceedings of the Croatian parliament in its 9th term (2016-2020), 20.65 million words in size. The human transcripts were normalized via a rule-based normalization method[3] inspired by (Ebden and Sproat, 2015), primarily focused on expanding numerals, acronyms and abbreviations.

The audio modality of the parliamentary proceedings was collected from the official YouTube channel of the Croatian Parliament[4] where all video recordings of the parliamentary debates are available. The recordings were downloaded with the youtube-dl tool[5] and encoded as wav in 16Khz, 16-bit, single channel.

The total length of the downloaded 755 audio recordings is 2821 hours, with an average length of 3.7 hours. The maximum length of a video is 6.76 hours. After applying Voice Activity Detection (VAD) (Silero, 2021) over the recordings, the length of pure speech recordings identified is 2,419.19 hours.

2.1. Initial Data Alignment via the Commercial ASR System

Given that no openly available training data or ASR system existed before the activities we describe in this paper, in the first iteration of our efforts, we had to rely on commercially available ASR systems. We chose to use the Google STT system due to the fact that it supports the Croatian language, and that it was supported in the code base used for this initial alignment, based on simple forced alignment(Plüss et al., 2020), which assumes that both data modalities come in the same, monotonic order. This required some rough manual alignment to be performed first, making sure that each identified subsection of audio and transcripts is in monotonic order. Given that we have in the meantime developed an alignment method which does not require monotonic data ordering, described in Section 2.3, this manual step will not be needed in the future application of our bootstrapping approach. Using the speech-to-text Google ASR system proved to fit into the 300 hours of Google Cloud usage that comes for free, so no fees were paid to Google.

From all the (audio, human transcript, automatic transcript) triplets generated in the process, 96 hours in length, we decided to keep those where the Levenshtein distance between the two transcripts, normalized by the average length of these transcripts, is equal or lower than 0.2. Loosely speaking, this means that only those segments were kept where the two transcripts are not different more than 20% on the character level. We identified this threshold to be useful via manual inspection. The resulting initial ASR training dataset was 72 hours in size.

2.2. Full Data Alignment via the In-House ASR System

With the initial ASR training dataset, consisting of 72 hours of speech recordings and normalized human transcript, we trained our in-house ASR system that would allow us to automatically transcribe, and then align, all the data available from the ParlaMint corpus and the Croatian parliament's recordings collection. Our chosen ASR technology was the recently released transformer-based multilingual XLS-R model (Babu et al., 2021), which showed to provide very good transcription results already with limited amount of training data. XLS-R is a multilingual model that was pretrained also on Croatian raw speech data.

We split the available data into a 66-hours training portion and 3-hour development and testing portions, running the fine-tuning procedure for 8 epochs. The preliminary evaluation results of the fine-tuned model over the initial training dataset are 13.68% of word error rate (WER) and 4.56% of character error rate (CER). This ASR system was used to transcribe the whole collection of audio recordings of 2,419.19 hours. With this we were able to drop our small collection of 72 hours of transcribed data and focus solely to producing the new collection, which also included most, if not all, of the initially aligned data.

Once the whole collection of audio recordings was successfully transcribed, we moved to the non-trivial problem of aligning human transcriptions available in the ParlaMint corpus and the automatic transcriptions obtained from our initial ASR system.

This alignment was hard in particular because of a different ordering of utterances in the recordings and in the official transcripts that are part of the ParlaMint corpus. It was due to the attempt by the transcribers to achieve logical (thematic) grouping instead of chronological ordering in the transcripts. Because of that, the explanation of the bill, reports of the parliamentary bodies dealing with the bill, discussions by the parliamentary groups representatives, and MPs themselves,

[2] https://cloud.google.com/speech-to-text/

[3] https://github.com/danijel3/TextNormalize

[4] https://www.youtube.com/c/InternetTVHrvatskogasabora

[5] https://github.com/ytdl-org/youtube-dl/

were followed by possible voting on amendments, and voting on the bill itself, although they may have taken place over the span of several days or weeks.

The process of obtaining the best utterance-level alignment of the whole corpus involved several steps. The desired output was supposed to contain short utterance segments (no longer than 20 seconds) with matching transcripts. The transcripts are matched on the audio level, so the pronunciation of all the words is naturally assumed, but due to imperfections of the simple rule-based normalization system, a match with the original, unnormalized transcript was also required. Furthermore, not all audio was transcribed in the original text and those fragments that were transcribed could contain simplifications, abstractions, deletions (e.g. due to substandard or unintelligible speech) or simply errors. It was decided that in this initial step, not all the data need be included and the rest can be completed once a better ASR system is developed, or by performing manual verification.

After obtaining the automatic transcription using our in-house system described above, the next step was to match the VAD derived segments to the human transcript without knowing their location or order within the larger transcript. The technical details of the procedure are described in section 2.3.

Once the match between the audio and human transcripts was acquired, there were still many errors that mostly occur on the boundaries of segments - especially if those boundaries are internal to continuous speech (i.e. not neighboring silence). To mitigate this further, the segments were joined together to longer portions of speech (up to 20 minutes long) and this was then aligned using the standard Viterbi forced alignment by our alternate WFST ASR system. This gave us word-level alignment of the whole corpus, from which the desired 20 second segments were easily extracted.

Following the above procedure, we managed to process around 82% of the input data which gave us a total of 1,976.97 hours of aligned speech with matching 14M words of human transcription.

2.3. Matching ASR Output to Long Text Corpora

This section provides a description of an engineering problem and what is undoubtedly just one of many possible solutions for it, but was optimal for our case. Ultimately, the problem can be described as matching a sequence of short text segments to a large corpus. There are sub-sequences of the shorter texts that occur in the same order in the large corpus, but all those matches have a level of discrepancy due to both errors in the ASR output as well as within the human transcripts.

To begin with, the whole text (both short and long) was converted to integer sequences where each word is represented by a single number instead of character strings - this improves both space and time complexity of the rest of the process. To start things off, we need to perform global lookup of the beginning of each sub-sequence within the long text. We do this by looking at all the locations of the first word within the first segment of the sub-sequence and choosing the position that has the smallest word-level Levenshtein edit distance. If no match is found or the match is not unique, we repeat the same process with the next word or segment and simply offset the result.

Once we identify the start of the sub-sequence, we heuristically match the rest of the segments by iterating forward until the match falls below a certain threshold. Then we start by repeating the global lookup procedure again and treat the rest of the segments as a new subsequence. If this also fails, we simply exit the procedure and discard the rest of the file.

The code of our approach will be released at the time of the final version of this manuscript.

3. Dataset Description and Availability

The full data alignment procedure resulted in 1,976.97 hours of speech recordings and their respective human and automatic transcripts. We applied the same filtering criterion via the normalized Levenshtein distance between human and automatic transcripts as with our initial dataset, discarding all the alignments where more than $\sim$20% of the transcripts differ. Applying this filter removed 146.58 hours, and thus we have obtained our final dataset presented in this paper, consisting of 1,816.34 hours of spoken data and their transcription.

We decided to brand the resulting dataset under the name ParlaSpeech-HR, encoding thereby our efforts as a continuation of the ParlaMint project, as well as the hope that many ParlaMint corpora will become ParlaSpeech datasets in the near future. This is especially crucial for a number of low-resource languages where parliamentary data are quite likely the single best source of a significant amount of spoken data and human transcripts.

The ParlaSpeech-HR corpus consists of 403,925 entries, each of which consists of the following attributes: (1) a path to the wave file, which also represents the ID of the entry, (2) the name of the original YouTube file, (3) the start (in milliseconds) of the entry in the original recording, (4) the end (in milliseconds) of the entry in the original recording, (5) a list of words from the human transcript, (6) local time offsets for each word in the human transcript, (7) a list of words from the normalized human transcript, (8) local time offsets for each word in the normalized human transcript, and (9) manually corrected normalized words, available for 484 entries only, used in an analysis in Section 4, (10) speaker information, if the segment was produced by only one speaker. Out of all 403,925 entries, only 22,076 entries have multiple speakers present, where we omitted the speaker information for simplicity. Each speaker description consists of the following information: (1) name, (2) gender, (3) year of birth, (4) party affiliation, (5) party status (ruling coali-

tion or opposition).

Overall there are 310 speakers present in the dataset, the most prominent one having 21,761 instances, the least frequent one having only two. Out of the 310 speakers, 234 are men, while 76 are women. There are 317,882 instances spoken by men, and 63,967 instances spoken by women.

To be able to use the dataset for benchmarking purposes, the dataset was further divided in a training, a development and a test portion, with three goals in mind: (1) having as many diverse speakers in the test portion, (2) having a gender balance in the test portion, and (3) not wasting unique speaker information on the development set. Having these three goals in mind, we decided to proceed as follows. Development data consist of 500 segments coming from the 5 most frequent speakers (four men and one woman), while test data consist of 513 segments that come from 3 male (258 segments) and 3 female speakers (255 segments). There are no segments coming from the 6 test speakers in the two remaining subsets. Given that there are 22,076 instances without speaker information, and therefore not being assigned to any of the three subsets, the training subset consists of the remaining 380,836 instances. The assignment of each of the instances into the three subsets (train, dev, test) is encoded as the last piece of information in the description of each instance. Segment-level statistics from the final dataset are presented in Table 1. We make the dataset freely available under the CC-BY-SA license via the CLARIN.SI repository.[6]

4. Correspondence of the Manual Transcripts and the Audio Recordings

Given that parliamentary transcripts are regularly a standardised approximation of what was actually said, we performed a short analysis of those differences by manually correcting 484 segments consisting of 2.16 hours of speech.

Comparing the automatically normalized segments with their manually corrected counterparts, we obtain a WER metric value of 4.69% and a CER metric of 2.64%, pointing towards the conclusion that there was a rather low level of interventions necessary in the transcripts, but also that a ceiling for any automatic evaluation lies at these two measurements. Manual interventions are present in 290 segments, i.e., 59.9% of all analysed segments, showing that this low noise is still rather distributed across all segments.

Manually inspecting a subset of the differences showed the issues to be mostly due to inconsistency in the work of transcribers (primarily regarding the compliance to the standard or elimination of fillers), typos introduced by transcribers, and in less frequent cases, issues with automatic normalization (either wrongly normalized phenomena or unnormalized phenomena).

[6] http://hdl.handle.net/11356/1494

It is important to remember that we filtered out 7% of all segments in which the automatic and manual transcriptions were in significant disagreement, and some of those segments probably also consisted of examples where the transcripts differed more significantly from what was actually uttered.

5. Experiments with the ParlaSpeech-HR Dataset

5.1. Training Initial Baseline Systems

Our initial systems were trained on the initial ASR dataset, containing 66 hours of speech, using either the Kaldi toolkit (Povey et al., 2011), or the HuggingFace library (Wolf et al., 2019). They Kaldi systems served two purposes: as a baseline to monitor progress of our transformer-based systems and as a fast alternative to perform speech-to-text alignment. Table 2 shows several results obtained using different models.

The first experiment used models pre-trained on a different, commercial dataset (McAuliffe et al., 2017), that was later fine-tuned on our training set in the second experiment. The next two experiments used the TDNN and chain model architectures commonly used in Kaldi (Povey et al., 2016). The final baseline system is the initial XLS-R-based model. The results show a significant improvement when the baseline model is further trained on our data, as well as the superior performance of transformer-based models.

5.2. Training ASR Systems on Normalized or Original Text

The two entries in the middle part of Table 2 compare XLS-R when trained on original and normalized transcripts, respectively. Given the significant capacity of transformer models, our hypothesis was that we could train future models on original data without need for noisy normalization of training data and de-normalization of automatically transcribed text. To reiterate, normalization was primarily focused on expanding numerical values in digital format and frequent acronyms and abbreviations.

In these experiments 110 hours were used for training.We trained the XLS-R transformer model for 8 epochs. The results in Table 2 show that there is barely any difference in the quality of the two outputs. Given that the normalized phenomena are not highly frequent, we performed a short focused analysis of the ASR output trained on the original text. The transformer model showed to be highly effective both on generating numerals in the digit form, as well on acronyms and abbreviations that occurred in the test dataset.

5.3. Comparing XLS-R and Slavic Models

Given the quick developments on the front of speech transformer models, during the finalisation of this paper, a new model pre-trained only on the Slavic por-

	sum	min	max	mean	median
spoken (seconds)	6,538,823	8	20	16.2	19.1
original (# of words)	14,533,541	1	82	35.7	38
normalized (# of words)	14,679,339	1	84	36.1	38

Table 1: Segment-level statistics calculated over the 403,925 segments available in the ParlaSpeech-HR dataset. Information is given on the spoken mode, and the original human and transcripts, and the automatically normalized transcripts.

System	WER	CER
GMM/WFST baseline	66.92%	50.43%
GMM/WFST adapted	30.54%	12.60%
TDNN/WFST	22.51%	9.78%
TDNN/WFST chain	16.38%	6.91%
XLS-R-66-initial	13.94%	5.42%
XLS-R-110-original	10.57%	3.23%
XLS-R-110-normalized	10.15%	3.04%
XLS-R-300	7.61%	2.34%
Slavic-300	6.79%	2.22%
Slavic-300+lm	4.30%	1.88%

Table 2: Output of various ASR systems. The first group of experiments was performed on the initial 66 hours of training data, the second on 110 hours, and the third on 300 hours.

tion of the VoxPopuli dataset has emerged,[7] which had been pre-trained on 89.9 thousand hours of raw speech. In this, last set of experiments, we compare the XLS-R model to the Slavic model, investigating whether a model pre-trained on a narrower set of languages, but no new data, would perform better. We fine-tune each of the models on 300 hours of training data. The experiments are performed on the original data, so no normalization of the transcripts was performed.

The results presented in the first two rows of the third section of Table 2 show that the Slavic model seems to be slightly better than XLS-R, with a relative error reduction of 11% on WER and a 5% error reduction on CER.

Given that the HuggingFace transformers library recently included support for adding language models to the ASR process, we perform a final experiment with adding to the Slavic model a 5-gram language model, trained on the whole ParlaMint corpus. The performance of the model further improves to 4.3% of word-error-rate and 1.88% of character-error-rate. While these results might sound very strong, the relatedness of the training, the testing, and the language model data has to be taken into account, and further experiments are needed on more diverse data.

We release the three models described in this section in the HuggingFace model repository.[8]

6. Conclusion

With the development of the ParlaSpeech-HR dataset we believe to have put the Croatian language on the map of the hastily developing language technologies. During our experiments we have shown that (1) automatic alignment of non-monotonic speech and human transcripts is very much possible, (2) there is significant difference in the spoken content and the human transcripts of parliamentary proceedings (around 5% of words are affected), (3) transformer models significantly outperform Kaldi-based models, (4) for transformer models it is not important for the data to be normalized as these have enough capacity to learn to produce digits and frequent abbreviations, (5) models pre-trained on a more limited set of related languages seem to perform better than general multilingual models, and (6) even in the case of data where the training and the testing domains are similar, a language model can still improve the output of transformer models.

We will continue our efforts by (1) producing the multimodal corpus of Croatian parliamentary proceedings, (2) performing speaker profiling experiments, and (3) applying the presented bootstrapping methodology to other under-resourced languages in dire need of similar datasets. We have high hopes that the ParlaSpeech methodology is a great opportunity for other under-resourced languages to obtain cheap, high-quality ASR datasets. Along these hopes, we make our code available at `https://github.com/clarinsi/parlaspeech`.

Acknowledgements

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

This work was also funded by the CLARIN ERIC project "ParlaMint", the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N) and the research programme "Language resources and technologies for Slovene" (P6-0411).

[7] `https://huggingface.co/facebook/wav2vec2-large-slavic-voxpopuli-v2`

[8] `https://huggingface.co/classla/wav2vec2-xls-r-parlaspeech-hr`
`https://huggingface.co/classla/wav2vec2-large-slavic-parlaspeech-hr`
`https://huggingface.co/classla/wav2vec2-large-slavic-parlaspeech-hr-lm`

7. Bibliographical References

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., et al. (2021). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Ebden, P. and Sproat, R. (2015). The Kestrel TTS text normalization system. *Natural Language Engineering*, 21(3):333–353.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Grigorova, V., Rudolf, M., Pančur, A., Kopp, M., Barkarson, S., Steingrímsson, S., van der Pol, H., Depoorter, G., de Does, J., Jongejan, B., Haltrup Hansen, D., Navarretta, C., Calzada Pérez, M., de Macedo, L. D., van Heusden, R., Marx, M., Çöltekin, Ç., Coole, M., Agnoloni, T., Frontini, F., Montemagni, S., Quochi, V., Venturi, G., Ruisi, M., Marchetti, C., Battistoni, R., Sebők, M., Ring, O., Dargis, R., Utka, A., Petkevičius, M., Briedienė, M., Krilavičius, T., Morkevičius, V., Diwersy, S., Luxardo, G., and Rayson, P. (2021). Multilingual comparable corpora of parliamentary debates ParlaMint 2.1. Slovenian language resource repository CLARIN.SI.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., et al. (2022). The ParlaMint corpora of parliamentary proceedings. *Language resources and evaluation*, pages 1–34.

Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., and Gudhnason, J. (2017). Building an ASR Corpus Using Althingi's Parliamentary Speeches. In *INTERSPEECH*, pages 2163–2167.

Katsamanis, A., Black, M., Georgiou, P. G., Goldstein, L., and Narayanan, S. (2011). SailAlign: Robust long speech-text alignment. In *Proc. of workshop on new tools and methods for very-large scale phonetics research*.

Kirkedal, A., Stepanović, M., and Plank, B. (2020). FT speech: Danish parliament speech corpus. In *Interspeech 2020*. ISCA, oct.

Mansikkaniemi, A., Smit, P., Kurimo, M., et al. (2017). Automatic Construction of the Finnish Parliament Speech Corpus. In *INTERSPEECH*, volume 8, pages 3762–3766.

Marasek, K., Koržinek, D., and Brocki, Ł. (2014). System for automatic transcription of sessions of the Polish senate. *Archives of Acoustics*, 39(4):501–509.

Martinčić-Ipšić, S., Ribarić, S., and Ipšić, I. (2008). Acoustic modelling for Croatian speech recognition and synthesis. *Informatica*, 19(2):227–254.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech*, volume 2017, pages 498–502.

Nouza, J., Safarik, R., and Cerva, P. (2016). ASR for South Slavic Languages Developed in Almost Automated Way. In *INTERSPEECH*, pages 3868–3872.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Plüss, M., Neukom, L., Scheller, C., and Vogel, M. (2020). Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German text corpus.

Popović, B., Ostrogonac, S., Pakoci, E., Jakovljević, N., and Delić, V. (2015). Deep neural network based continuous speech recognition for Serbian using the Kaldi toolkit. In *International Conference on Speech and Computer*, pages 186–192. Springer.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Interspeech*, pages 2751–2755.

Silero. (2021). Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. `https://github.com/snakers4/silero-vad`.

Solberg, P. E. and Ortiz, P. (2022). The Norwegian Parliamentary Speech Corpus. *CoRR*, abs/2201.10881.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. 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 993–1003, Online, August. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.