Improved Open Source Automatic Subtitling for Lecture Videos

Robert Geislinger[1,2] Benjamin Milde[1,2] Chris Biemann[1]

[1]Language Technology Group, Universität Hamburg, Germany
[2]Hamburger Informatik Technologie-Center e.V., Germany
robert.geislinger@uni-hamburg.de
benjamin.milde@uni-hamburg.de
christian.biemann@uni-hamburg.de

Abstract

This paper summarizes the current state of development in improving an open source subtitling tool. This includes improvements to the speech recognition model for German, the replacement for the punctuation reconstruction architecture and the addition of an audio segmentation. The goal of these adjustments is an overall better subtitle quality. The most crucial part of the existing pipeline, the German speech recognition, is replaced by a new Kaldi TDNN-HMM model trained on 70% of additional audio data, resulting in a word error rate of 6.9% on Tuda-De. The punctuation reconstruction model for German texts is replaced by a Transformer-based approach that is also trained on new data. English is added as a fully supported second language, including speech recognition and punctuation reconstruction models. Furthermore, to improve speech recognition in long videos, audio segmentation was also added into the pipeline to support long videos flawlessly without quality issues.

1 Introduction

Remote learning with lecture videos has become the norm in the Covid-19 pandemic. Subtitling videos make them accessible for persons with hearing limitations. Since subtitling videos by hand is a time-consuming and cost-intensive task, this work offers a solution for automatic subtitling. Automatic speech recognition (ASR) is the most important step in the creation of subtitles, but for sufficient results, the text must also be supplemented with punctuation marks and be separated at appropriate places to achieve a good reading flow.

This paper presents the results of a revised pipeline to create German and English subtitles with open source algorithms and models. It also introduces the addition of audio segmentation as well as improvements to automatic speech recognition and punctuation reconstruction models. The entire pipeline is shown in Figure 1. The model for German ASR was revised and a model for English language was added. Also, the existing punctuation reconstruction model is replaced by a new Transformer-based architecture and trained on new data. It is now also possible to get live status information about the current processing step via a Redis database.

The tool is already in operation at the Universität Hamburg lecture video portal Lecture2Go[1] and the generated subtitles serve as a starting point for further manual annotation. Users of the platform can also correct the subtitles with a web-based subtitle editor.

2 Related Work

Generating subtitles with ASR can be performed both semi-automatically and automatically. In semi-automatic generation systems, texts are respoken in a controlled environment by a trained speaker (Sperber et al., 2013; Romero-Fresco, 2020; Vashistha et al., 2017). However, automatic systems are already being used to subtitle videos and conferences (Milde et al., 2021; Geislinger et al., 2021).

There are several models for German speech recognition available. A model based on Kaldi TDNN-HMM with ARPA rescoring and RNNLM achieved a word error rate (WER) of 7.4% on Tuda-De (Milde, 2022). The currently lowest WER on Tuda-De is a Conformer Transducer model with 5.8%, which is trained on about 4,600 hours of training data (Wirth and Peinl, 2022). The model presented in this paper with Kaldi TDNN-HMM architecture is trained on about 1,720 hours with a WER of 6.9%. A model for English speech recog-

[1] https://lecture2go.uni-hamburg.de

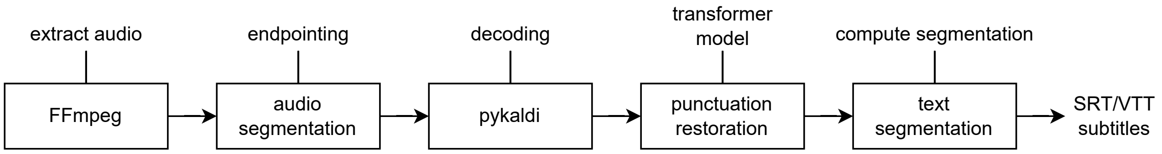

Figure 1: Full processing pipeline of the tool

nition achieved a WER of 5.9% on Switchboard (Tüske et al., 2021).

For punctuation reconstruction, there are also several available models. Multilingual models for German, English, French as well as Transformer based models for Polish (Chordia, 2021; Guhr et al., 2021; Wróbel and Zhylko, 2021). Recurrent Neural Networks are used by Hládek et al. (2019) to supplement a Slovak speech recognition system.

3 Speech Recognition Models

The most important feature that is needed in order to create suitable and understandable subtitles is a well-trained ASR model. This work is divided into the improvement of an existing, freely available German model speech recognition model and the creation of a new English speech recognition model under the Apache License 2.0 . Kaldi was used as a speech recognition framework to train our ASR models, as it is under the Apache License 2.0 and provides multiple training scripts for German and English, which were used as a starting point for this work (Povey et al., 2011). For decoding, we use Kaldi's nnet3 lattice decoder with PyKaldi (Can et al., 2018).

3.1 German Model

For automatic speech recognition in German, the freely available Kaldi-Tuda-De model was used as a basis for improvement. The training script uses 1,000 hours of audio data to train the acoustic model and about 100 million German sentences from several free available sources to train the language model (Milde and Köhn, 2018).

The training data for the acoustic model was increased from 1,000 hours by 720 hours to a total of 1,720 hours. This was achieved by replacing the Common Voice version 3 data set with the updated Common Voice version 8 data set (Ardila et al., 2020). This resulted in an expansion of the number of speakers about all data used from 5,546 to a total of 16,929. One of the model training data sets is Tuda-De which was also revised in this work to remove errors (Radeck-Arneth et al., 2015). Several broken audio files in the test and training data were removed and corrections were made to the transcript. In total, these corrections removed less than one minute of data, which is far less than one percent of the total data.

The training data for the language model were also part of the revision with the aim to achieve a lower WER and also to incorporate current words and terms into the language model. The data was crawled for this purpose from several freely available sources with the german-asr-lm-tools[2] project. The data consist mainly of articles from the news program Tagesschau, German Wikipedia, subtitles of German TV stations such as ARD and proceedings of the EU Parliament (Koehn, 2005).

The script to train the model itself was also improved to remove pitfalls in the training and make it easier to train and extend it with additional data for individual purposes (e.g. adding university lectures as training data). This should also give persons with limited language processing knowledge the possibility to train a model for their requirements.

The modifications in the Tuda-De data set and the additional data for the language and acoustic model lead to lower WER. The previous WER of the model was 14.4% with a lexicon of more than 350,000 words and without LM rescoring (Milde and Köhn, 2018). The newly trained model lowered the WER to 10.2% which is 29% relatively lower. This may be due to the increased lexicon of more than 900,000 words as well as the 70% more data.

When also using ARPA and RNNLM rescoring the model performs at 6.9% WER which is a relative reduction of 52% compared with the previous model. The results in comparison with other models are shown in Table 1. The training script and pretrained models are available[3] under the Apache License 2.0.

[2] https://github.com/bmilde/german-asr-lm-tools/

[3] https://github.com/uhh-lt/kaldi-tuda-de

System	Model	Data	test WER
Radeck-Arneth et al., 2015	TDNN-HMM hybrid, FST	108h	20.5
Milde and Köhn, 2018	”	375h	14.4
Milde, 2022	”	1720h	7.4
Wirth and Peinl, 2022	E2E / Conformer CTC	4520h	7.8
”	E2E / Conformer T	”	5.8
This model	TDNN-HMM hybrid, FST	1720h	6.9

Table 1: The WER results of the German models on the Tuda-De test set

3.2 English Model

To support speech recognition for English videos as well, an own expandable training script for English was created. The script is based on the TEDLIUM TDNN-HMM script for Kaldi. The TEDLIUM corpus consists of recordings of TED Talks. In total, the data set contains 118 hours of audio data (Hernandez et al., 2018).

To expand the training data, the Librispeech corpus was added. Librispeech contains recordings of audiobooks of the LibriVox and Gutenberg Project (Panayotov et al., 2015). This dataset is read speech, i.e. books read aloud in a quiet environment. A total of 100 hours of audio data are added to the script. This makes a total training data for the acoustic model of 218 hours.

Language model training material was expanded by YouTube subtitles from the pile data set. These additional texts add current topics and words to the training data (Gao et al., 2020). To prepare the texts, punctuation as well as languages other than English are removed. The toolkit to clean up English texts for language modelling in an ASR contest is available as a separate project[4]. Unknown words in the lexicon were added by using a Sequitur G2P model (Bisani and Ney, 2008), which was trained on already existing words in the combined lexicon of the TEDLIUM and Librispeech data set.

After Arpa and RNNLM rescoring the WER of the new model is 13.1% on Librispeech test set ”test-other” and 4.8% on ”test-clean” which is 12% lower compared to the model by Panayotov et al., 2015. On the TEDLIUM test data the WER is 10.3% which is 53% higher than the model by Hernandez et al., 2018. In their current state, the results on the TEDLIUM test set are still clearly in need of improvement. This can be achieved by adding further data sets like Gigaspeech, increasing the training data for the language model or train on further adapted training scripts (Chen et al., 2021). The results are shown in Table 2. The training script and pretrained English ASR models are available[5] under the Apache License 2.0.

System	Data	WER	
		LS	TED
Panayotov et al., 2015	100h	5.5	
Hernandez et al., 2018	118h		**6.7**
This model	218h	**4.8**	10.3

Table 2: The WER results of Kaldi TDNN-HMM models on librispeech and TEDLIUM test set

4 Punctuation reconstruction

Text transcriptions generated by ASR often lack punctuation and capitalization. To make the text more human-readable in post-processing, punctuation is reconstructed. For German punctuation reconstruction, Milde et al. (2021) used Punctuator2 which was trained on 5 million lines of German text. This architecture is based on a recurrent neural network (Tilk and Alumäe, 2016). The goal of this work is to outperform the error rate of the German model and also train an English model. For both languages, pretrained BERT-based models are used. As a starting point to fine-tune the models, the trainings scripts of Daulet Nurmanbetov[6] are used. The pretrained German model used for later fine-tuning is GBERT (Chan et al., 2020). The German punctuation reconstruction model is fine-tuned on 94 million lines of German subtitles and Wikipedia articles. For evaluation, the NoSta-D corpus was used (Benikova et al., 2014). The model by Milde et al., 2021 achieved an error rate

[4] https://github.com/uhh-lt/english-asr-lm-tools

[5] https://github.com/uhh-lt/kaldi-asr-english

[6] https://github.com/Felflare/rpunct

Model	System	error rate
Milde et al., 2021	BRNN	9.1%
This model	BERT-based	6.2%

Table 3: Comparison German Punctuation reconstruction error rates on NoSta-D for period, comma and questionmark

of 9.1% for reconstruction of period, comma and question mark in German texts. The new model achieved an error rate of 6.2% which is relative reduction of 31%. The results are also shown in Table 3.

5 Changes in the Tool Pipeline

Further changes to the pipeline involve an added language selection, audio segmentation and process feedback. The pipeline with all parts is shown in Figure 1. The language can now be changed before each video and the languages are managed via a configuration file. To support a wider range of Kaldi models, support for CMVN and RNNLM rescoring was added to the decoder.

5.1 Audio segmentation

Processing longer videos as a whole can lead to unpredictable behavior in Kaldi. This can result in segments being skipped and gaps in the transcript. One reason for this behavior is the rising memory demand with every minute of decoding. To work around this problem and process videos of several hours running time flawlessly, the file must be split into smaller chunks. The easiest approach could be a hard cut after a fixed amount of time but that would also cut in the middle of words and thus increase the error rate. To avoid the problem of splitting during a word, an beam search based endpointing algorithm was implemented (Reddy, 1976).

The algorithm finds the best segmentation that breaks on pauses in the signal. It also seeks to fulfill an average segment length criteria (default 1 minute). For this, the energy of the signal is analyzed and splitting costs are assigned to all positions in the audio. The energy function is smoothed with a Gaussian filter, so that longer periods of low energy (longer pauses) have the lowest splitting cost. The search algorithm combines this with a segment length criteria and finds a solution that compromises between both criteria. These resulting segments can be passed to Kaldi as input. This also makes it possible for later enhancements to use multithreading to maximize the performance of the pipeline by decoding the segments simultaneously.

5.2 Process feedback

The new version of the tool adds also additional functionality to receive update messages about the progress of the pipeline when using the tool in a backend (e.g. a video platform). The tool sends information to registered services via a Redis pub/sub channel. These messages contain information about the current processing step. The status messages can be used to visualize the progress to a frontend while creating the subtitles. The additional feedback helps the user to understand the current progress of the processing job and there is also more information should a processing step fail.

6 Conclusion

Creating automatic subtitles for videos needs a lot of well-tuned models to attain good results. Even if an ASR system is the most important part of the pipeline, good models for punctuation reconstruction are also a necessity for well readable subtitles. Previously, our tool was only able to subtitle German videos. We were able to improve the German ASR model and significantly improved WER results. We also expanded language support and added models for English. Further additions presented in this paper added more possibilities in the existing tool, especially when used in a backend of a video platform.

The subtitling software is published[7] under the Apache License 2.0, with instructions and download scripts for all necessary models.

7 Outlook

Since the project is still in development at this point, we hope that the results will continue to improve. This concerns in particular the punctuation model as well as the English ASR model.

When Kaldi's successor K2 (Żelasko et al., 2021) is more stable, a new German and English model based on the presented training scripts can be developed and trained. With this new architecture and additional data sets, this could also lead to better results due to new acoustic modelling techniques.

The reconstruction of punctuation could be further optimized with usage of Transformer-based

[7] https://github.com/uhh-lt/subtitle2go

models. This could be done with more training data and also with new models and architectures. Platforms with Transformer models bring a wide range of pre-trained models and training scripts (Wolf et al., 2019). Research on the post-processing pipeline could also lead to a new end-to-end model to summarize the different steps into one specially adapted model for the purpose of subtitle creation. Besides the added English models, other languages could bring the project to a wider audience outside of German and English videos.

For longer videos, multithreading could be used on the segmented audio, to transcribe different parts of one video in parallel.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. Nosta-d named entity annotation for german: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland.

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.

Dogan Can, Victor Martinez, Pavlos Papadopoulos, and Shrikanth Narayanan. 2018. PyKaldi: A python wrapper for Kaldi. In *Proc. Acoustics, Speech and Signal Processing (ICASSP)*, pages 5889–5893, Calgary, Canada.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online).

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech 2021*, pages 3670–3674, Brno, Czech Republic.

Varnith Chordia. 2021. PunKtuator: A multilingual punctuation restoration system for spoken and written text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 312–320, Online. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Robert Geislinger, Benjamin Milde, Timo Baumann, and Chris Biemann. 2021. Live Subtitling for BigBlueButton with Open-Source Software. In *Proc. Interspeech 2021*, pages 3319–3320, Brno, Czech Republic.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2021. Fullstop: Multilingual deep models for punctuation prediction. In *Proceedings of the Swiss Text Analytics Conference 2021*, Winterthur, Switzerland.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *International conference on speech and computer*, pages 198–208.

Daniel Hládek, Ján Staš, and Stanislav Ondáš. 2019. Comparison of recurrent neural networks for slovak punctuation restoration. In *2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 95–100, Naples, Italy.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Benjamin Milde. 2022. *On Representation Learning in Speech Processing and Automatic Speech Recognition*. Ph.D. thesis, Universität Hamburg, Germany.

Benjamin Milde, Robert Geislinger, Irina Lindt, and Timo Baumann. 2021. Open source automatic lecture subtitling. In *Proceedings of ESSV 2021*, pages 128–134, Virtual Berlin, Germany.

Benjamin Milde and Arne Köhn. 2018. Open source automatic speech recognition for German. In *Proceedings of ITG 2018*, pages 251–255, Oldenburg, Germany.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 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, Brisbane, Australia.

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*, Hilton Waikoloa Village, Hawaii, USA.

Stephan Radeck-Arneth, Benjamin Milde, Arvid Lange, Evandro Gouvea, Stefan Radomski, Max Mühlhäuser, and Chris Biemann. 2015. Open Source German Distant Speech Recognition: Corpus and Acoustic Model. In *Proceedings Text, Speech and Dialogue (TSD)*, pages 480–488, Pilsen, Czech Republic.

Raj Reddy. 1976. *Summary of Results of the Five-Year Research Effort at Carnegie-Mellon University*. Carnegie-Mellon University, Department of Computer Science.

Pablo Romero-Fresco. 2020. *Subtitling through speech recognition: Respeaking*. Routledge.

Matthias Sperber, Graham Neubig, Christian Fügen, Satoshi Nakamura, and Alex Waibel. 2013. Efficient speech transcription through respeaking. In *Proceedings of Interspeech 2013*, pages 1087–1091, Lyon, France.

Ottokar Tilk and Tanel Alumäe. 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Proceedings of Interspeech 2016*, pages 3047–3051, San Francisco, California, USA.

Zoltán Tüske, George Saon, and Brian Kingsbury. 2021. On the limit of english conversational speech recognition. *arXiv preprint arXiv:2105.00982*.

Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2017. Respeak: A voice-based, crowd-powered speech transcription system. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 1855–1866, Denver, Colorado, USA.

Johannes Wirth and Rene Peinl. 2022. Asr in german: A detailed error analysis. *arXiv preprint arXiv:2204.05617*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv e-prints*, pages arXiv–1910.

Krzysztof Wróbel and Dmytro Zhylko. 2021. Punctuation restoration with transformers. In *Proceedings of the PolEval 2021 Workshop*, pages 33–37, Warsaw, Poland.

Piotr Żelasko, Daniel Povey, and Sanjeev Khudanpur. 2021. Speech recognition with next-generation kaldi (k2, lhotse, icefall). In *Proc. Interspeech 2021*, Brno, Czech Republic.