*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6363–6367
Marseille, 11–16 May 2020

# Large Corpus of Czech Parliament Plenary Hearings

**Jonáš Kratochvíl, Peter Polák, Ondřej Bojar**
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
surname@ufal.mff.cuni.cz

## Abstract

We present a large corpus of Czech parliament plenary sessions. The corpus consists of approximately 444 hours of speech data and corresponding text transcriptions. The whole corpus is segmented to short audio snippets making it suitable for both training and evaluation of automatic speech recognition (ASR) systems. The source language of the corpus is Czech, which makes it a valuable resource for future research as only a few public datasets are available for the Czech language.

We complement the data release with experiments of two baseline ASR systems trained on the presented data: the more traditional approach implemented in the Kaldi ASR toolkit which combines hidden Markov models and deep neural networks and a modern ASR architecture implemented in Jasper toolkit which uses deep neural networks in an end-to-end fashion.

**Keywords:** Czech, automatic speech recognition, dataset, speech corpus

## 1. Introduction

The field of automatic speech recognition (ASR) has been recently undergoing a methodological shift from hybrid model architectures towards end-to-end systems. These mostly neural-network-based architectures have recently been gaining momentum in their popularity and obtained state-of-the-art results on multiple popular speech corpora, e.g. Han et al. (2019) on test-clean set from LibriSpeech (Panayotov et al., 2015).

One particular downside of the end-to-end approaches is their requirement of an extensive amount of training data in order to produce competitive results in comparison with more traditional hybrid architectures.

At the same time, there has been active research in the field of fluent speech reconstruction (Fitzgerald et al., 2009; Cho et al., 2016; Klejch et al., 2017), which aims to produce a formal textual report of a particular recording. In reality, speakers often repeat certain words multiple times, correct themselves in the middle of a sentence or use informal language to describe their ideas. In many applications such as meeting reports, automatic subtitling of presentations or subsequent machine translation of the ASR output, we would like the transcripts to be a formal-language equivalent of what was said and not the exact copy of the actually uttered words.

In our paper, we present an extensive collection of Czech parliament plenary sessions which tries to contribute to both speech recognition per se as well as speech reconstruction. The corpus consists of 444 hours of spoken Czech together with the corresponding formal transcriptions as obtained by stenographers present at the meetings.

Our paper is organized as follows. In Section 2., we describe related literature and relevant work. In Section 3., we introduce our methodology used when constructing the corpus. Section 4. contains a detailed corpus description and its exploratory analysis. Section 5. presents the baseline results we obtained and we discuss them in Section 6. After a brief overview of our future plans (Section 7.), we conclude the paper in Section 8.

## 2. Related Work

There are relatively few publicly available speech-related resources for the Czech language.

The largest available Czech speech corpus is Prague DaTabase of Spoken Czech, PDTSC (Hajič et al., 2017)[1]. It comprises of approximately 122 hours of spontaneous dialog speech. The corpus is interesting not only for its size but also because it contains three layers of transcriptions: automatic speech recognition output aligned to audio (denoted as *z-layer*), literal manual transcript (*w-layer*) and finally speech reconstruction (*m-layer*).

There already exists a corpus of Czech parliament hearings published by the University of West Bohemia (Pražák and Šmídl, 2012), who collected the data from the year 2011 between February and August. The corpus comprises of 88 hours of transcribed speech. Another related Czech speech corpus is Vystadial 2016 (Plátek et al., 2016) which is a dataset comprising of 78 hours of telephone conversations in Czech.

With the same thematic as our corpus, there are some publicly available datasets for other languages. Multilingual dataset Europarl-ST (Iranzo-Sánchez et al., 2019) is a corpus of European Parliament debates in six languages. The most prominent parliament hearing corpus is a 2000-hours corpus of Finnish parliament (Mansikkaniemi et al., 2017). Another speech corpus is Althingi's Parliamentary Speeches of Icelandic parliament (Helgadóttir et al., 2017). There is also a relatively large corpus of 249 hours of Parliament hearings collected in Bulgarian parliament as described in Geneva et al. (2019).

## 3. Methodology

Our source data are available at the website of Chamber of Deputies[2] and they come in 15-minute audio segments. The corresponding texts for each of these audio files do not

[1] http://ufal.mff.cuni.cz/pdtsc1.0/

[2] http://www.psp.cz/eknih/2017ps/audio/2017/index.htm

match the exact spoken words as there is always some overlap between adjacent audio recordings. For this reason, we have first downloaded and concatenated all audio files for a particular hearing, removing the overlaps.

We also semi-automatically pre-processed the corresponding text transcriptions for individual parliament hearings. This pre-processing step includes mainly the removal of the text parts. We remove texts that do not correspond to the spoken audio. We also removed commentaries of some events during the hearing included by the stenographer. We normalized numerical values to text strings ("33" to "thirty three") and removed all non-speech related parts of the transcriptions.

The resulting concatenated audio files for each hearing are between 1 and 12 hours long. For segmentation of these long audio files to smaller parts, we used the Kaldi speech recognition toolkit (Povey et al., 2011). First, we trained a TDNN model (Peddinti et al., 2015) using our in-house Czech speech data and later used this model as the base model for the Kaldi audio segmentation script.[3] This method subdivides the text into shorter documents and performs decoding of the audio with a language model strongly biased towards the current document of interest. Note that these segments are related to the sound signal, not to sentence boundaries in the transcript.

After the segmentation step, we cleaned the data by discarding segments that did not match the decoded output. We have run this process in two iterations, always training a new model on the most recently segmented clean data and using it to generate new segments of the original files. The resulting audio segments are between 1 second and 44 seconds long.

## 4. Corpus Description

In this section, we provide an overview of the presented corpus. First, the way how the Czech parliament operates is described, then the obtained collection of the data, and finally, the corpus composition as we release it for the research community.

### 4.1. Czech Parliament

Czech parliament provides an audio recording of all their hearings together with the corresponding stenographic transcriptions starting from the year 2017. This period spans one election term of the Czech parliament.

During each parliament session, speakers usually take turns in communicating their opinions about a particular topic that is on the agenda. Speakers can also react to each other and ask for further clarification. In most cases, the hearing has a relatively predictable structure. First, the list of absent parliament members is read, and the proposed changes in the agenda are discussed and voted on. Second, the topic of the hearing is introduced, and members of the parliament take turns in describing their opinions and reacting to each other. Lastly, voting about the proposed changes takes place, and the hearing moves to the next topic on the agenda.

| | |
|---|---:|
| Audio hours | 444 |
| Audio files | 191455 |
| Number of hearings | 85 |
| Longest audio segment | 44s |
| Shortest audio segment | 0.5s |
| Unique speakers | 212 |
| Female speakers | 48 |
| Male speakers | 164 |
| Total words | 3 029 646 |
| Unique words | 221 638 |
| Time period | Nov. 2017 - Nov. 2019 |

Table 1: Corpus statistics

### 4.2. Corpus Statistics

In total, we have collected 85 parliament hearings over the considered period. These hearings are between 1 and 12 hours long.

The text transcriptions come from professional stenographers, who also revise the transcriptions and remove duplicate or repeating words, rewrite informal parts of speech to their formal equivalent and add non-speech marks, such as *laughter*, *background noise*, or a short description of the current situation in the parliament if it is noteworthy for the complete record of the hearing. During the plenary hearing, the stenographer on duty is seated directly next to the current speaker. There are between 8-10 stenographers present during each plenary hearing. Stenographers take turns every 10 minutes in transcribing the speech. Between their transcription turns, they polish the texts in order to adhere to the unified style specified by the Czech parliament.

We estimate that there is about 1-5% mismatch between the actual words the speaker said and the text that occurs in the transcript. We also note that the positive semantic correlation between the mismatch transcriptions and the actual words is very strong. The mismatch parts are, in most cases, more formally rewritten parts of the actual speech, which was too informally or colloquial.

We present the ASR corpus statistics in Table 1.

### 4.3. Final Corpus Composition

The final corpus consists of three versions of the data:

**ASR Segments** The audio is segmented to short parts, each complemented with the corresponding transcript in uppercase and without punctuation.[4] This version is directly usable for the training of the ASR acoustic model or end-to-end ASR.

**ASR Transcript** The plain text version of the full hearings, in the form suitable for ASR, i.e., uppercased, no punctuation, numbers spelled out, etc.

**Plaintext Transcript** This version of the transcript is sentence-oriented, true cased, and it also includes speaker identifiers, transcriber commentaries, and time information.

[3] https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/steps/cleanup/segment_long_utterances_nnet3.sh

[4] If the full-length unsegmented recordings are required, please contact authors.

| Section | Hours | Words |
|---|---|---|
| Training | 433.5 | 2 966 148 |
| Development | 3.0 | 20 595 |
| Test | 7.5 | 42 903 |

Table 2: Sections of the released corpus.

The combination of ASR transcript and Plaintext transcript is useful for the training of sentence segmentation models. The direct output of an ASR system on the sound files together with plaintext transcript is useful for the training of speech reconstruction systems.

All three corpus versions were split into training, development, and test sets in the same way. The development set was taken to be a held-out part of the training set, whereas the test set was taken from a parliament hearing in a different election term.

The development set thus matches very closely the statistical distributions of the training data (but it, of course, contains unseen speech segments). The test set is deliberately made more challenging, avoiding speaker and topic overlap with the training data as much as possible.

We present the respective training, development and test set statistics in Table 2.

## 5. Experiments

In this section, we describe experiments with the proposed corpus of Czech Parliament Plenary Hearings. We conduct two experiments: first, we use the more traditional approach implemented in the Kaldi toolkit which uses the combination of Hidden Markov model (HMM), together with the Gaussian Mixture Model (GMM) and deep neural network. Second, we experiment with the end-to-end deep neural ASR architecture Jasper (Li et al., 2019).

We provide the results of our experiments in terms of word error rate (WER).

### 5.1. Kaldi

Kaldi (Povey et al., 2011) has been a popular speech recognition toolkit, especially for languages with limited data resources. The training of the model is done in two stages. First, we train the HMM-GMM model and make the alignments between speech and transcript of our training data. We then train a neural network on top of these alignments, which makes the acoustic model more robust. Kaldi is based on the underlying weighted finite-state transducers, which allows flexible incorporation of the n-gram language model that we use in our experiments as well.

#### 5.1.1. Architecture Overview

For audio feature representation, we use 13 MFCC features (Davis and Mermelstein, 1980) together with their first and second-order derivatives, therefore our inputs for both the HMM-GMM and neural network training are 39-dimensional vectors. We first train the GMM-HMM model and make data alignments. As the neural network acoustic model, we use a combination of convolutional and time-delayed layer (Peddinti et al., 2015) neural network. For language modeling, we use the 4-gram language model, in the KenLM implementation (Heafield et al., 2013) trained on the corpus transcriptions.

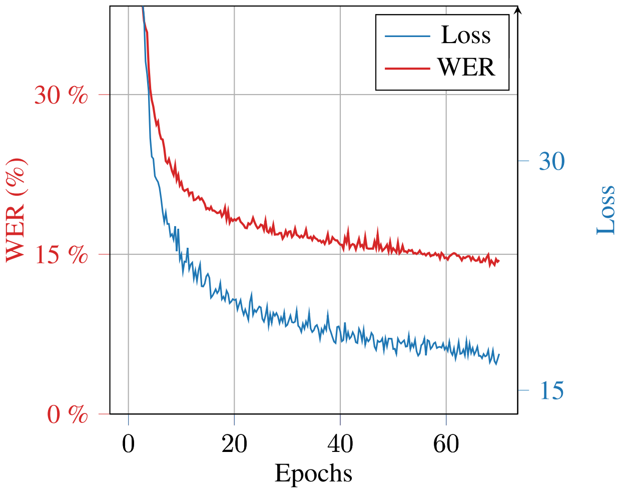

Figure 1: Jasper learning curve: the performance on the development set in terms of training loss and word error rate

### 5.2. CNN Jasper

Jasper is a family of end-to-end, deep convolutional neural network ASR architectures that unite acoustic and pronunciation models within one model. We decided to experiment with Jasper as it is a good example of contemporary end-to-end ASR solutions and because of its availability within the OpenSeq2Seq toolkit (Kuchaiev et al., 2018).

#### 5.2.1. Architecture Overview

The input of the model are mel-filterbank features from 20 ms windows with 10 ms overlap. The output of the model is a probability distribution over characters from a custom vocabulary.

Jasper applies one pre-processing and three post-processing convolutional layers. Between these layers is the main part of the network, which consists of so-called “blocks”.

The main part of the Jasper model consists of $B$ blocks and $R$ sub-blocks (the authors introduced convention where each Jasper model can be described as “Jasper $B$x$R$”). In our experiment, we use Jasper 10x5.

A sub-block comprises a 1D-convolution, batch normalization, ReLU activation, and finally, dropout. Each block input is connected to the last sub-block via residual connections. 1x1 convolution is used in order to project input channels to a different number of output channels. After each convolution layer, batch normalization is applied. The batch norm output is added to the output of the batch norm layer in the last sub-block. The sum is passed to the activation and dropout and produces the output of the current block.

Residual connections inspired by DenseNet (Huang et al., 2017) are employed to enable such a deep network to converge.

#### 5.2.2. Training

As previously stated, we use the OpenSeq2Seq toolkit to build and train our end-to-end ASR model. Specifically, we

| Method | Decoding | Dev WER | Test WER |
|---|---|---|---|
| Jasper | Greedy decoding | 11.59 % | 14.24 % |
| | Beam search | 10.57 % | 14.25 % |
| | Beam search LM | 9.09 % | 9.93 % |
| Kaldi | Beam search LM | 6.64 % | 7.10 % |

Table 3: Experimental results of Kaldi a Jasper models on development and test sets.

make use of an existing implementation and configuration as provided by Jasper authors directly. Its main advantages include reduced training time and the implementation of mixed-precision training (all weights are stored in float16 while weights updates are computed in float32).

We altered the output vocabulary and extended it with characters used in the Czech language (characters with acute, caron, ring). Because of limited time and resources, we trained our model only for 70 epochs (instead of 400 as in the original paper). According to OpenSeq2Seq toolkit documentation (NVIDIA, 2018), authors indicate that training for 50 epochs should still yield acceptable results (for LibriSpeech dev-clean set it ought to be under 5% WER versus 3.61% after 400 epochs). We keep other network parameters unchanged.

We trained the network for three days and 20 hours on 8 Quadro P5000 GPUs. The progress and quality of the training were continuously checked on the development set, see Figure 1.

### 5.3. Results

We have used 433.5 hours of the training data to performs the experiments and evaluate the models on the development and test set.

The development set was randomly extracted from the training data, so the speaker's independence between the training and development set is not explicitly ensured. (It is possible but unlikely that a rare speaker was exclusively selected for the development set.)

The test set was especially chosen with the intention to reduce the possibility of speaker overlap with the training data, to arrive at a more realistic setting. With unseen speakers in the test set, any overfitting to the training data would be discovered.

We compare three methods of decoding: greedy decoding (for each time frame, the most probable character is selected), beam search and beam search with language model rescoring (Hannun et al., 2014).

In this experiment, we use the 4-gram language model in the KenLM (Heafield et al., 2013) implementation trained on the transcripts of the training set. Both, the plain beam search and the beam search witch language model rescoring use beam width of 256. The results can be seen in Table 3.

## 6. Discussion

From the results, we see that the Kaldi model slightly outperforms Jasper architecture. This may be mainly caused by the fact that end-to-end systems require even larger training data in order to match or improve over the hybrid architectures.

We have also used only n-gram language models in our experiments. In the future, we would like to add Transformer (Vaswani et al., 2017) as a language model, in a separate rescoring phase. This could substantially improve the results because the Transformer can capture longer contextual information than n-gram models.

## 7. Future Work

In the future, we want to process additional recordings and transcription from Czech parliament and senate using our existing data preparation pipeline.

We believe that we can extract an additional 1000 hours of audio data together with their transcriptions, which are already available. Moreover, as the parliament hearings take place each month, and there is a constant inflow of data at the rate of approximately 30 hours per month, so we would like to automate the whole dataset preparation pipeline and extend our dataset on the fly with each new hearing release.

Further, we would like to refine segmentation and alignment using the Jasper ASR model pre-trained on the current dataset. Al will allow us to segment recordings with respect to sentences rather than voice activity detection boundaries, enabling to train e.g., end-to-end ASR models with a built-in language model.

Another challenge of our interest is to balance the dataset with respect to the individual speakers (e.g., the president speaker opens every session) and also allow ensure gender balance, if desired.

## 8. Conclusion

We presented a new Czech speech corpus of 444 hours of Czech parliament plenary hearings. Further, we ran baseline experiments with speech recognition systems trained on the new corpus. We demonstrate that the proposed corpus is suitable for training both the traditional ASR models such as Kaldi, but also the state of the art end-to-end neural networks, yielding excellent results on benchmark development and test sets that we have prepared.

We release the corpus online for public use at:

`http://hdl.handle.net/11234/1-3126`

## 9. Acknowledgements

This research work was supported by the the project no. 19-26934X (NEUREM3) of the Czech Science Foundation and ELITR (H2020-ICT-2018-2-825460) of the EU research and innovation program.

## 10. Bibliographical References

Cho, E., Niehues, J., Ha, T.-L., and Waibel, A. (2016). Multilingual disfluency removal using nmt. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT), Seattle, USA*.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

Fitzgerald, E., Jelinek, F., and Frank, R. (2009). What lies beneath: Semantic and syntactic analysis of manually reconstructed spontaneous speech. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 746–754, Suntec, Singapore, August. Association for Computational Linguistics.

Geneva, D., Shopov, G., and Mihov, S. (2019). Building an asr corpus based on bulgarian parliament speeches. In Carlos Martín-Vide, et al., editors, *Statistical Language and Speech Processing*, pages 188–197, Cham. Springer International Publishing.

Hajič, J., Pajas, P., Ircing, P., Romportl, J., Peterek, N., Spousta, M., Mikulová, M., Grůber, M., and Legát, M. (2017). Prague DaTabase of spoken czech 1.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Han, K. J., Prieto, R., Wu, K., and Ma, T. (2019). State-of-the-art speech recognition using multi-stream self-attention with dilated 1d convolutions. *arXiv preprint arXiv:1910.00716*.

Hannun, A. Y., Maas, A. L., Jurafsky, D., and Ng, A. Y. (2014). First-pass large vocabulary continuous speech recognition using bi-directional recurrent dnns. *arXiv preprint arXiv:1408.2873*.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., and Guðnason, J. (2017). Building an asr corpus using althingi's parliamentary speeches. In *INTERSPEECH*, pages 2163–2167.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Iranzo-Sánchez, J., Silvestre-Cerdà, J. A., Jorge, J., Roselló, N., Giménez, A., Sanchis, A., Civera, J., and Juan, A. (2019). Europarl-st: A multilingual corpus for speech translation of parliamentary debates. *arXiv preprint arXiv:1911.03167*.

Klejch, O., Bell, P., and Renals, S. (2017). Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5700–5704. IEEE.

Kuchaiev, O., Ginsburg, B., Gitman, I., Lavrukhin, V., Li, J., Nguyen, H., Case, C., and Micikevicius, P. (2018). Mixed-precision training for nlp and speech recognition with openseq2seq.

Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J. M., Nguyen, H., and Gadde, R. T. (2019). Jasper: An end-to-end convolutional neural acoustic model. *arXiv preprint arXiv:1904.03288*.

Mansikkaniemi, A., Smit, P., Kurimo, M., et al. (2017). Automatic construction of the finnish parliament speech corpus. In *INTERSPEECH*, pages 3762–3766.

NVIDIA. (2018). Jasper — openseq2seq 0.2 documentation.

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.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Plátek, O., Dušek, O., and Jurčíček, F. (2016). Vystadial 2016 – czech data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Pražák, A. and Šmídl, L. (2012). Czech parliament meetings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*.