Proceedings of the Federated Conference on Computer Science and Information Systems pp. 185–188
DOI: 10.15439/2020F119
ISSN 2300-5963 ACSIS, Vol. 21

Czech parliament meeting recordings as ASR training data

Jan Oldřich Krůza
Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics,
Charles University
Email: kruza@ufal.mff.cuni.cz

***Abstract*—I present a way to leverage the stenographed recordings of the Czech parliament meetings for purposes of training a speech-to-text system. The article presents a method for scraping the data, acquiring word-level alignment and selecting reliable parts of the imprecise transcript. Finally, I present an ASR system trained on these and other data.**

I. Introduction

TRAINING data for speech recognition is always a demanded commodity, especially if it is free. There are for sure already some free Czech corpora fit for speech recognition training:

- Vystadial[1] with its 77 hours of VoIP calls[2],
- The Prague Database of Spoken Czech[3] with its 122 hours of richly annotated spontaneous dialogues[4],
- The Czech Senior COMPANION Expressive Speech Corpus with its 5 hours of professionally spoken utterances by a single speaker[5],
- Otázky Václava Moravce: 35 hours of transcribed recordings of the Czech TV talk show[6],
- STAZKA, a set of speech recording from vehicles with its 35 hours of background noise and utterances[7],
- Spoken Corpus of Karel Makoň[8] with its 100 hours of manually transcribed spontaneous speech by a single speaker[9],
- and possibly others that I am not aware of.

The Czech parliament meeting recordings represent a publicly available dataset of high-quality audio recordings of contemporary Czech in consistent low-noise audio quality worth almost 4000 hours of downloadable material, about 2800 hours after subtraction of the overlaps. Extracting training data for speech recognition systems would provide a corpus at least one order greater in length than those so far publicly available.

Verily, I am not the first person to attempt using these recordings for speech recognition. The Department of Cybernetics of University of West Bohemia developed an automatic online subtitling system for the meetings in 2006[10] and as a result, an 88-hour subset annotated by high-quality automatic transcript has been released for speech recognition training purposes[11].

I attempt to use the official stenographic transcripts available for all the talks so that it can be a new entry in the above list, on par in quality and excelling in size.

II. Data Preparation

Since the source data is publicly available and in the public domain, I merely provide the scripts for downloading and building the corpus. The algorithms and parameters used are described in this section.

A. Scraping

Regrettably, the data are to my best knowledge only available in human-readable form. The transcript is not clearly distinguished in the markup and is interlaced with metainformation. My method of isolating the transcript is quite crude but it covers the vast majority of cases. The criterion is to extract the subtree of all nodes with HTML attribute `[align=justify]`, except HTML elements `<b>`, which contain speaker identification.

The known shortcomings of this method are that 1) it discards the speaker annotations, although it is valuable metainformation and 2) it skips some short passages, e.g. references to other meetings, as can be seen in the meeting from Feb. 12th 2020 10:10 - 10:20[1]. Both can be corrected by devising a smarter scraper and neither has any significance for speech recognition: speaker annotation fundamentally and neglecting the links for their infrequency.

B. Alignment

One of the obstacles in using the stenographic transcripts for training an ASR system is the very loose alignment available. The recordings are all 14 minutes long and have a 4-minute overlap. The corresponding transcript is thus aligned in 10-minute blocks with a roughly 2-minute padding on each side of the audio. Figure 1 schematically shows the alignment of the stenographic transcript to the audio and the overlap of the recordings.

Systems for aligning long audio segments to their transcripts already exist, like that of Moreno et al.[12] or Hazen[13]. They are both based on an already existing automatically acquired transcript. I use this technique as well, though simplified and adapted to the task.

I have used the dataset mentioned above[11] to train a GMM-based ASR system, using the stenographs as training

[1]https://www.psp.cz/eknih/2017ps/stenprot/040schuz/s040372.htm

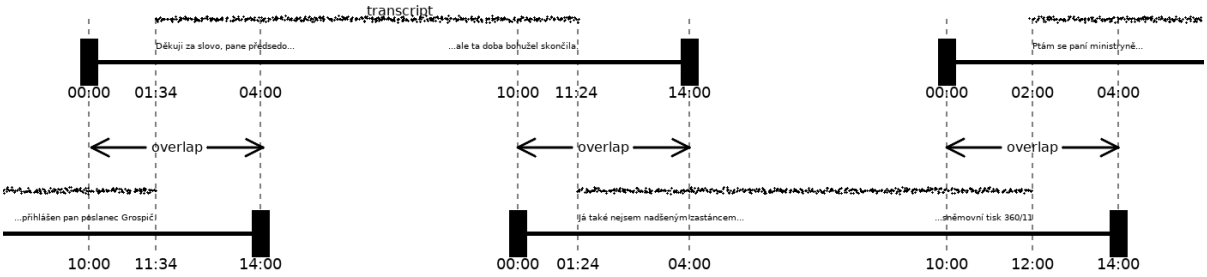

Fig. 1. Alignment and overlap of audio files and transcript. The examples are from Feb. 12th 2020 around 10 o'clock. The transcript corresponding to the recording in the upper left covers audio positions 01:34 - 11:24. The one in the lower right from 01:24 to 12:00.

data for a language model. Using these models, a word-level-aligned transcript of the whole set of recordings has been acquired.

The predicted transcript and the stenographic one have then been compared for Levenshtein distance, determining the edit operations needed to transform one into the other. For each predicted word, a reliability score is then computed as 1 - unreliability where unreliability is the number of edit operations taken on it divided by its length. Figure 2 shows how the stenographic transcript is aligned with the audio on word level.

Nota bene, a GMM-based system was chosen for the initial transcript instead of a DNN-based for three reasons: 1) Foremost, it is straightforward to obtain precise alignment from a GMM-based system. 2) The training doesn't require so much computational resources and data. 3) It isn't crucial to have maximum possible accuracy in this stage.

C. Audio Segmentation

To create a usable dataset for training a speech-to-text system, it is not necessary to perfectly align the whole transcript. On the contrary, it is desirable to align what is reliably precise and discard the rest.

The criteria for good training samples are:

1) 100% precise transcript,
2) roughly sentence-level length,
3) consistent length.

To ensure precise transcript, it is good to have the samples padded by some silence, since the alignment obtained from the initial ASR may be a bit imprecise. We thus want to split at pauses, the longer the better, up to a certain limit (about 1 second). The need to split at longer pauses goes against the need to split at consistent, none-too-great lengths.

So the problem is to select an optimal set of silences so that the longest ones are used and so that they split the recording into chunks of length in a given range. This looks like a problem for dynamic programming but a simpler approach is also possible: Start with a set of all silences predicted by the forced alignment. Iterate over the silences shortest-first and remove each if it doesn't break the constraints.

I have experimentally set the length boundaries to 12 - 30 seconds. The maximum length could be decreased at the cost of available pauses to choose from, which would lead to more frequent splits in the middle of a word.

D. Training Samples Selection

With the audio segmented and corresponding manual transcripts extracted, the last step remaining is selecting which segments to include in the traning data. Indeed, since the recordings have a 2-minute padding on each side for 10 middle minutes, we must discard at the very least 40% of the segments. I use the following criteria for including a segment in the data:

1) The first and last token have reliability at least 70%,
2) The mean reliability of all tokens is at least 70%,
3) The number of words is no less than five.

Minimum reliability of border tokens is considered to minimize the danger of shifted alignment boundaries. Mean reliability is considered because it is OK for some words to have very low reliability: there are enough errors in the prediction, that's why we use the manual transcript after all. But if too many tokens have too low reliability, then it is a sign of a suspicious segment. The number of words has a minimum because with only a few words, the probability of misalignment with good score is much greater than when there are enough words.

Why use mean reliability and not median? The way the reliability is computed considers the number of edit operations on one line in the automatic transcript. In the case where there are many insertions, the reliability of one line can go arbitrarily deep sub zero. So it can happen that there are several inserted words in a (mis)aligned chunk that only affect the reliability score of a single word. The mean taps these while the median doesn't.

E. Data Extraction Summary

All the constants and criteria are to be considered a baseline solution. They all could be tweaked much more rigorously and solved much more soundly. However, this simple solution readily yields a high-quality training dataset of 1058 hours. Of the total 539,057 segments, 142,530 (26%) have been accepted

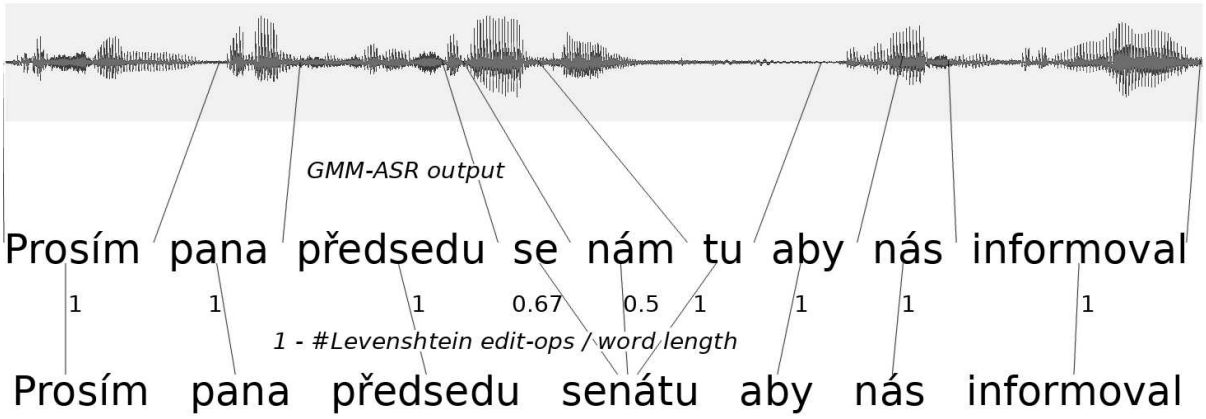

Fig. 2. Schema of aligning the audio to the stenographic transcript on word level.

to the training dataset. Of the total 396,527 discarded segments, 350,258 (88%) were discarded because of the criterion of unreliable start or end. It should be noted however, that the start / end reliability criterion is applied first, so it catches segments that would be discarded for other reasons also.

Reducing the minimum reliability of the boundary words from 70% to 50% increases the number of accepted chunks by 17%. It adds 5% segments of the total number to the dataset. But if we consider that 40% of the total number of segments must be discarded because of audio padding, the gain is acually 9%. It is an option to increase the training data volume at the cost of matching precision.

III. Numerals and Abbreviations

There are many numeral expressions in the transcripts. They amount to 489,880 out of 25,010,269 tokens in the complete stenographic transcript, which is almost two percent. In the training dataset, 24% of the samples contain one or more numerals.

Originally, I have included the digits into the alphabet for speech recognition, thus attempting to train the system to transcribe numeral expressions directly into digits. The speech recognition system described in the following section would however transcribe numeral expressions as empty strings.

There are four ways to deal with the problem:

1) ignore it,
2) remove digits from the training data,
3) manually expand digits to words,
4) automatically expand digits to words.

The first option needs no elaboration. The second one, removing samples with digits, is an easy and viable option but it is a waste of a quarter of the dataset and of the vast majority of samples with numerals in them. Manual expansion would surely be ideal but very costly. It remains to attempt the fourth variant of automated expansion.

For automated expansion of digits into words, we can use the available initial transcript and the algorithm for alignment with the stenographic transcript.

The expansion is done in two steps:

1) generation of verbal variants,
2) selection of the most likely variant.

I have used the Perl module `Lingua::CS::Num2Word` as a base for the expansion. I modified the module in the following way: 1) I added support for the order of billions, which is very common in the corpus. 2) A number is no longer expanded into a single phrase but instead into all possible phrases expressing the given number. 3) I added support for genitive and accusative cases, decimal numerals, ordinals, dates and times.

All tokens in the stenographs that include digits are expanded into their verbalization variants before further processing. Upon alignment, the variant with least edit distance from the initial transcript is selected.

Common abbreviations and symbols are expanded together with the digits. For example, the very common character *"§" (paragraph)* is expanded into the forms *paragraf, paragrafu, paragrafů, paragrafem, paragrafech* that represent common inflections of the word. Some common abbreviations that undergo inflection include *"čl." (article)*, *"odst." (also paragraph)* and *"tzv." (co-called)*.

After incorporating the expansion into the pipeline, the similarity of the stenographic transcript and the initial one raised, which also raised the number of accepted segments from 26% to 35%. The amount of training data grew by 86 hours to 1144.

IV. ASR Based on the Dataset

I have trained a standard DeepSpeech[14] model on the 1058 hours with training : development : test ratio of 18 : 1 : 1; batch size 50; learning rate 0.0001; dropout rate 0.2. The training took 12 epochs to reach optimal dev fit and the final word error rate on testing data from the corpus itself is 8.40% before digit expansion and 7.89% afterwards.

The language model used was a pentagram model with pruned singleton trigrams, tetragrams and pentagrams. The

bulk of scraped transcriptions, including those with no downloadable corresponding audio, was used as training data for the language model.

I have also tried training a speech recognition system with other datasets and the combination of them all. Of the datasets listed in section I, only Vystadial, Otázky Václava Moravce (ovm) and the corpus of Karel Makoň (makon) proved useful without much effort.

Apart from them, I used the publicly not available corpora of Charles University Corpus of Financial News (CUCFN, 65 hours)[15], the Balanced corpus of informal spoken Czech (Oral2013, 293 hours)[16] and the spoken Bible (100 hours) available with no license terms from poslouchamebibli.cz. Table I shows the speech recognition results for each corpus on test data from itself and on a common test set from all the corpora.

TABLE I
WORD ERROR RATE OF SPEECH RECOGNITION ON THE INDIVIDUAL CORPORA AND ON THEIR CONCATENATION.

source	WER on self	WER on all
bible	9.20%	94.7%
cucfn	31.6%	72.8%
makon	30.4%	77.3%
oral2013	78.4%	60.7%
ovm	21.6%	72.9%
parliament w/digits	8.74%	39.7%
parliament expanded	7.89%	**36.0%**
vystadial	51.0%	74.0%
all w/digits	28.4%	28.4%
all expanded	26.0%	26.0%

All speech recognition systems were trained with the same hyperparameters as described above.

V. CONCLUSION

I have presented a new corpus of spoken Czech suitable for training speech recognition systems based on data in the public domain. The corpus size exceeds by an order the size of other freely available such corpora. A speech recognition system with competitive performance was made to show the fitness of the dataset to the purpose.

Among the compared corpora, the Czech parliament corpus performs by far best even in speech recognition outside its domain.

Source code for scraping and building the corpus is in the public domain and available on GitHub.com/Sixtease/cz-parliament-speech-corpus.

ACKNOWLEDGMENTS

This work has been using language resources developed, stored and distributed by the LINDAT/CLARIAH project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

This research was supported by SVV project number 260 575.

REFERENCES

[1] M. Korvas, O. Plátek, O. Dušek, L. Žilka, and F. Jurčíček, "Free english and czech telephone speech corpus," 2014.

[2] O. Plátek, O. Dušek, and F. Jurčíček, "Vystadial 2016 – czech data," 2016, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. [Online]. Available: http://hdl.handle.net/11234/1-1740

[3] M. Mikulová, J. Mírovský, A. Nedoluzhko, P. Pajas, J. Štěpánek, and J. Hajič, "Pdtsc 2.0-spoken corpus with rich multi-layer structural annotation," in *International Conference on Text, Speech, and Dialogue*. Springer, 2017, pp. 129–137.

[4] J. Hajič, P. Pajas, P. Ircing, J. Romportl, N. Peterek, M. Spousta, M. Mikulová, M. Grůber, and M. Legát, "Prague DaTabase of spoken czech 1.0," 2017, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. [Online]. Available: http://hdl.handle.net/11234/1-2375

[5] M. Grůber, "Czech senior COMPANION expressive speech corpus," 2014, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. [Online]. Available: http://hdl.handle.net/11858/00-097C-0000-0023-1D76-9

[6] L. Šmídl and A. Pražák, "OVM – otázky václava moravce," 2013, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. [Online]. Available: http://hdl.handle.net/11858/00-097C-0000-000D-EC98-3

[7] L. Šmídl, P. Stanislav, and V. Radová, "STAZKA – speech recordings from vehicles," 2015, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. [Online]. Available: http://hdl.handle.net/11234/1-1510

[8] O. Krůza and N. Peterek, "Making community and asr join forces in web environment," in *International Conference on Text, Speech and Dialogue*. Springer, 2012, pp. 415–421.

[9] O. Krůza, "Spoken corpus of karel makoň," 2012, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. [Online]. Available: http://hdl.handle.net/11372/LRT-1455

[10] A. Pražák, J. V. Psutka, J. Hoidekr, J. Kanis, L. Müller, and J. Psutka, "Automatic online subtitling of the czech parliament meetings," in *International Conference on Text, Speech and Dialogue*. Springer, 2006, pp. 501–508.

[11] A. Pražák and L. Šmídl, "Czech parliament meetings," 2012, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. [Online]. Available: http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4

[12] P. J. Moreno, C. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Fifth International Conference on Spoken Language Processing*, 1998.

[13] T. J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Ninth International Conference on Spoken Language Processing*, 2006.

[14] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[15] W. Byrne, J. Hajič, P. Ircing, F. Jelinek, S. Khudanpur, J. McDonough, N. Peterek, and J. Psutka, "Large vocabulary speech recognition for read and broadcast czech," in *International Workshop on Text, Speech and Dialogue*. Springer, 1999, pp. 235–240.

[16] L. Benešová, M. Křen, and M. Waclawičová, "Korpus spontánní mluvené češtiny oral2013," *Časopis pro moderní filologii (Journal for Modern Philology)*, vol. 1, no. 97, pp. 42–50, 2015.