Towards a Speech Recognizer for Komi, an Endangered and Low-Resource Uralic Language

Nils Hjortnæs
Department of Linguistics
Indiana University
Bloomington, IN
nhjortn@iu.edu

Niko Partanen
University of Helsinki
Helsinki, Finland
niko.partanen@helsinki.fi

Michael Rießler
University of Eastern Finland
Joensuu, Finland
michael.riessler@uef.fi

Francis M. Tyers
Department of Linguistics
Indiana University
Bloomington, IN
ftyers@iu.edu

Abstract

In this paper, we present and evaluate a first pass speech recognition model for Komi, an endangered and low-resource Uralic language spoken in Russia. We compare a transfer learning approach from English with a baseline model trained from scratch using DeepSpeech (an end-to-end ASR model) and evaluate the impact of fine tuning a language model for correcting the output of the network. We also provides an overview of previous research and perform an error analysis with a focus on the language model and the challenges introduced by a fieldwork based corpus. Though we only achieve a 70.9% Character Error Rate, there is a great deal to be learned from the circumstances presented by our data's structure and origins.

1 Introduction

In the creation of any corpus of spoken text, the transcription work can be identified as the major bottleneck that limits how much recorded speech data can be annotated and included in the corpus. The situation is particularly dire with endangered languages for which language technology does not exist (Foley et al., 2018, 206). But typically, even corpus building projects working with spoken data from majority languages manage to transcribe and analyze only a fraction of the materials for which they have recorded audio data. The need for speech-to-text tools is not restricted to fieldwork-based language documentation producing new speech recordings, but rather a continuum of projects and languages with various levels of resources. There is also an immense build-up of non-transcribed legacy audio recordings of endangered languages stored at various private or institutional archives, in which case even a small and endangered language may have a significant amount of currently unused materials. At the same time, speech recognition technologies have been fully functional for a variety of languages for some time already. Although the use of such tools would potentially offer large improvements for language documentation and corpus building, it is still unclear how to integrate this technology into work with endangered languages in the most successful manner.

Spoken corpora of endangered languages for the study of endangered languages are often relatively small, especially when compared to the resources available for larger languages. This is not necessarily due to lack of relevant audio recordings. There are no statistics about the typical sizes of endangered language corpora, but it can be assumed that transcribed portions are somewhere from a few hours to tens of hours, with magnitudes of hundreds of hours becoming rare. This is much lower than the threshold usually estimated that is needed for major speech recognition systems. From this point of view, the initial goal of using speech recognition in this context could be attempting to improve the transcription speed. This would result in larger transcribed corpora which could continuously improve the speech recognition system. The accuracy needed to reach that point would be such that it is faster to correct than do transcription manually, as before then speech recognition doesn't help the tran-

Proceedings of the 6th International Workshop on Computational Linguistics of Uralic Languages, pages 31–37
Wien, Austria, January 10 — 11, 2020.
https://doi.org/10.18653/v1/FIXME

scription task.

2 Related work

There have been several earlier attempts to build pipelines that integrate speech recognition into language documentation context, most importantly Elpis (Foley et al., 2019) and Persephone (Adams et al., 2018). These systems are still maturing, with desire to make them more easily available for an ordinary linguist with no technical background in speech recognition. There are only individual reports of project having yet adapted these tools, with exceptions such as work described in Michaud et al. (2018) on the Na language, where an error rate of 17% was reported. Also Adams et al. (2018) report that it seems possible to achieve phoneme error rates below 30% with only half an hour of recordings. Both of these experiments were done in a single speaker setting.

Instead of using tools specifically designed in a language documentation context, in this paper we train and evaluate a speech recognition system for Zyrian Komi using DeepSpeech (Hannun et al., 2014).

DeepSpeech has been used previously with a variety of languages. It is most commonly used with large languages when the resources available vastly outnumber what we have. We found several other cases where DeepSpeech was used, for example, with Russian (Iakushkin et al., 2018), Romanian (Panaite et al., 2019), Tujian (Yu et al., 2019) and Bangla (Saurav et al., 2018). All of these experiments report higher scores than we do, with the exception of Russian, with smaller data, but there are important differences as well. Romanian recordings were done in studio environment, Tujian sentences were specifically translated to Chinese to take advantage of the Chinese model, and Bangla experiment had a limited vocabulary. The Russian corpus has well over 1000 hours, which brings it, in a way, out of the low-resource scenario where the other mentioned works took place.

One experiment with DeepSpeech that seems particularly relevant to us is the work done recently on Seneca (Jimerson et al., 2018) because the word error rate was very high and difficult to reduce.

The overview of related work leads us to the conclusion that speech recognition has reached significant results in conditions where very large transcribed datasets are available, or there are other constraints present, such as a small number of speakers and/or studio recording quality.

3 Komi language

Komi is a Uralic language spoken primarily in the North-Eastern corner of European Russia, bordering the Ural mountains in the East. There are, however, numerous settlements where Komi is spoken outside the main speaking areas, and these communities span from the Kola Peninsula to Western Siberia.

Zyrian Komi is closely related to Permian and Jazva Komi. All Komi varieties are mutually intelligible and form a complex dialect continuum. Komi is more distantly related to Udmurt, which is spoken south from main Komi areas. Together Komi and Udmurt form the Permic branch of Uralic languages. Other languages in this family are significantly more distantly related.

The Komi language currently has approximately 160,000 speakers, and it is spoken in a large number of individual settlements in Northern Russia. The language is taught, although to a limited degree, in schools as a subject in some municipalities. There are several weekly publications and the written language is stable and generally well known. There is also continuous online presence. The largest Komi corpus contains over 50 million words (Fu-Lab, 2019). For a more thorough description see, i.e. Hausenberg; Цыпанов (2009).

Komi is spoken in intensive contact with Russian, a dominant Slavic language of the region. A large portion of the Komi lexicon is borrowed from Russian, and virtually all speakers are currently bilingual. Bilingual phenomena present in contemporary Komi have been studied in detail (Leinonen, 2002, 2006), and with particularly importance for our study, the northern dialect that is predominantly present in our corpus is known for its extensive Russian contact (Leinonen, 2009).

Komi is written with Cyrillic orthography. The script is essentially phonemic, although different character combinations are used to represent similar sounds in different contexts, as is typical for Cyrillic scripts.

4 Resources used

4.1 The Spoken Komi Corpus

The majority of Komi resources used in this study originate from the Kone Foundation funded *Iźva Komi Documentation Project*, the results of which

are currently available in the Language Bank of Finland (Blokland et al., 2019). However, there are numerous Komi materials that are in various stages of being turned into corpora, and these include recordings stored in the Institute for the Languages of Finland. Eventually all these materials should be combined into the Spoken Komi Corpus, and developing speech recognition technologies that can operate on various recording types is an important part in advancing the work on these resources.

The corpus is relatively large, containing around 35 hours of transcribed utterances. The number of total recorded hours is much higher, as this count includes only the transcribed segments without silences. Also the number of individual speakers is very high, at over 200. This has been made possible by systematic inclusion of archival data, as the goal has been to build a corpus that is representative from different periods from which we have recordings, and also so that different geographical areas would be evenly covered.

Specific features of the corpus are that the majority of content consists of conversations between two or more native speakers. These conversations have been arranged in an interview-like setting, so one of the participant is leading the conversation with questions on various topics. The transcriptions are done by native Komi speakers, and have been systematically revised by one additional native speaking project participant besides the person who did the transcription. The recordings are very accurate in that small primary interjections such as 'mm' and 'aha' are transcribed. There is also a large amount of overlapping speech.

The transcriptions are in a Cyrillic writing system that follows the rules of Komi orthography. A similar system has been used in a recent Komi dialect dictionary (Безносикова et al., 2012). This convention was selected for various reasons, both practical and methodological. Having the results of language documentation work in written standard, when it exists, makes the work accessible for the community and allows better integration of language technology (Gerstenberger et al., 2017a,b). This is also obvious with the current study, as the speech recognition system that operates with the orthography is arguably more useful for the community than one which outputs a transcription system that only specialists in the field can easily understand. That being said, the use of orthography also makes some tasks such as speech recognition harder, as the phoneme-to-grapheme correspondence is less transparent.

Portion	Clips	Duration (Hours:Minutes)
train	37043	27:50
dev	4756	3:33
test	4736	3:28
Total:	46535	34:51

Table 1: Statistics on the training data

The texts in the corpus have been manually segmented into utterances and transcribed in ELAN. These segments have been transformed into pairs of audio and plain text files. For loading into DeepSpeech, the audio samples have been normalized for length such that clips over 10 seconds, DeepSpeech's default cutoff, are excluded.

4.2 DeepSpeech

DeepSpeech (Hannun et al., 2014) is a relatively simple Recurrent Neural Network designed specifically for the task of Speech Recognition. It has since been updated and made available[1]. The biggest change between the current 0.5.1 release of DeepSpeech and the original is the switch to an LSTM instead of an RNN. In addition, some hyperparameters have been updated. Unless otherwise noted, we use the default parameters in the 0.5.1 release.

Figure 1 outlines the structure of the DeepSpeech Neural Network. The feature extraction is a mapping of characters to the nominal values 1-N where N is the length of the set of characters appearing in the data. This is followed by three fully connected ReLU layers, the LSTM layer, and a final ReLU layer. All layers have a width of 2048. The sixth layer is a softmax layer with a width determined by the length of the alphabet.

The final step of DeepSpeech is correction using a language model (lm), which allows us to calculate the probability of a given character sequence. It is integrated into DeepSpeech by balancing the probability of the neural network's output with the probability of a character sequence in the lm (Hannun et al., 2014). The hyper-parameter alpha controls the degree to which the language model edits the neural network's output and the hyper-parameter beta controls the cost of inserting word breaks. A

[1]https://github.com/mozilla/DeepSpeech/releases/tag/v0.5.1

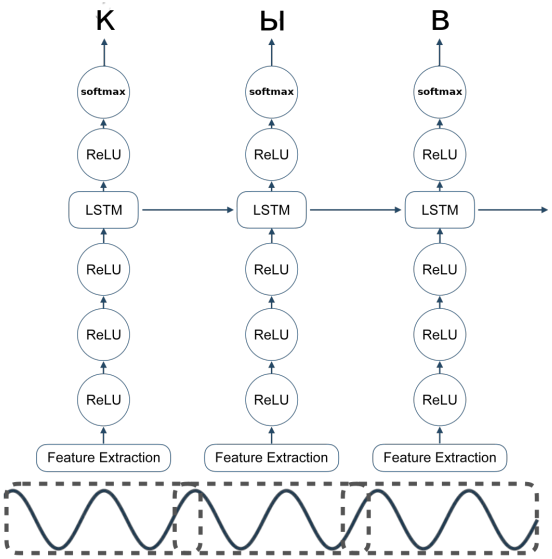

Figure 1: The architecture of Mozilla's DeepSpeech (Meyer, 2019)

higher alpha favors language model editing and a higher beta favors inserting word breaks.

5 Experiment

To pre-process the data, we shuffled it and split it into an 8-1-1 ration of training, testing, and development. We then created an alphabet of characters and symbols which appear in the text, the length of which determines the width of the output layer of the DeepSpeech neural network.

As a baseline, we trained DeepSpeech using the default parameters, except for batch sizes, on the Komi corpus from scratch. We then trained a transfer learning model on DeepSpeech, again with the default parameters except batch sizes, for comparison. Rather than using the default batch size of 1 for train, test, and dev, we used 128, 32, and 32 respectively for all experiments. Finally, we tuned the learning rate at factors of 10 from 0.001 to 0.000001 and dropouts of 5, 10, and 15%.

We trained the transfer learning model using the transfer_learning2[2] branch of DeepSpeech. This branch allows you to cut off the last N layers of the network and reinitialize them from scratch. This is necessary for the final layer because the alphabet, and therefore the width of the final output layer, will almost certainly change. Meyer (2019) found that cutting off two layers and transferring four when using DeepSpeech, as well as allowing fine-tuning of the transferred layers, provides the best boost in performance. We therefore follow suit, and cut off two layers and allow fine-tuning for our transfer models. For convenience, we used English as the source language because it ships with DeepSpeech and is known to have good results. Languages with comparable performance which are historically related to Komi, such as the main contact language Russian, provide potential avenues of research worth further experimentation.

A language model is a critical piece of DeepSpeech because it corrects for the fact that every character in the orthography is not pronounced in natural speech. We generated out n-gram trie language model, as in Hannun et al. (2014), using kenlm (Heafield, 2011) with the default parameters. Because a language model is trained on unlabeled text, we can train it on a much larger corpus than the speech dataset. Our corpus is composed of several books, newspaper articles, an old Wikipedia dump, and the Komi Republic website. These are all in the standard, modern Zyrian orthography. We found that the quantity of data provided by these various sources was more effective than using the transcriptions from our data.

Because the language model is applied to the output of the neural network, it can be tuned separately. Therefore, in the interest of time, we trained the network with the default language model hyperparameters of 0.75 and 1.85 for alpha and beta respectively. We then tuned the language model on the output from the best neural network for the baseline, transfer learning baseline, and tuned models. We tuned the lm for alphas of 0.25, 0.50, and 0.75 and betas of 1, 3, 5, 7, and 9, as can be seen in Tables 3 and 4.

In order to see whether the language model was helping or hindering our performance, we set both alpha and beta to 0, effectively disabling the influence of the language model entirely. This also allowed us to check the output of the neural network directly, as this also disabled the insertion of word breaks.

6 Results

The best results were achieved using the transfer learning model with a learning rate of 0.00001 and dropout of 10%. Early stopping was disabled as it is very aggressive, and all other parameters were the default or the batch sizes stated above as of release 0.5.1.

[2]https://github.com/mozilla/DeepSpeech/tree/transfer-learning2

System	CER (%)	WER (%)
Baseline	82.7	99.1
Transfer tuned	72.1	100.0
Transfer baseline	82.9	**98.3**
Baseline tuned lm	73.8	100.0
Baseline no lm	72.7	100.0
Transfer no lm	**70.9**	100.0

Table 2: The best results for our baseline and transfer learning models without tuning the language model, with tuning, and without a language model

Table 2 compares the best scores achieved for the baseline and transfer models under different conditions. The transfer models perform better under all respective conditions, but the baseline model outperforms the baseline transfer model when tuned. While tuning clearly has an effect on the Character Error Rate, the tuned models were unable to accurately recognize any full words. An error analysis showed that the words the baseline models were capturing are short filler words rather than content or even common function words. This is further discussed below.

Table 3 and Table 4 show the impact of the language model on the accuracy of the speech recognition system. A higher alpha favors correcting the output of the neural network with the language model, and a higher beta favors inserting word breaks. We see in both tables that a lower alpha achieves better results, corroborating Table 2, where disabling the language model achieved the best results. As alpha increases, the best results are achieved with increasing beta values as well.

7 Discussion

These preliminary results show that transfer learning is a promising avenue for developing a speech recognition system for documentary audio data. While the gain is small as compared to the baseline, any improvement in the network will help the language model better predict the true orthography. In addition, we found that the transfer model predicts slightly more sensible guesses than the baseline, even if it is not reflected in the error rates. For example, (1) and (3) are produced by the baseline and (2) and (4) are produced by the transfer model. Despite the overall error rate being high, both of these pairs of examples indicate that the potential for improvement is there, and that transfer learning is slightly more accurate.

(1) но печера ю вылын
н п зино юн

(2) но печера ю вылын
ино ече ю н

(3) но ме же том на
н м же м

(4) но ме же том на
н не же т

A negative indication of potential, however, is that several of the examples which are boosting the CER in particular are filler words such as *но*, *мм*, or *и*. That DeepSpeech is only good at identifying these exceptionally simple examples with a high degree of accuracy could be an indication of a class imbalance problem where the simple, small examples become too ingrained in the network and prevent more complex, more desirable behavior from emerging. For example, in (5), *но* and *и* appear in the output despite having no correlate in the source text.

(5) передовик вӧлэма
и ец теф и техо пняе но

DeepSpeech has built-in mechanisms for validating data before it is used, including skipping samples deemed too long or too short. For short audio clips, however, the threshold is fairly lenient. For this experiment, only two samples out of the 47232 were excluded for being too short. By increasing the minimum length of the audio clip for it to be valid, we can ignore these confounding data points and potentially improve the quality of the speech recognition.

Another way to refine the dataset would be to selectively choose data generated by certain speakers, such as those who contributed most to the corpus. As previously mentioned, there are over 200 speakers who have contributed to this corpus, but most of them are only a small portion. While this does decrease the potential for robustness when developing a generalized speech recognition system, it is less of an issue when considering the integration of speech recognition into field work and documentation, as there tend to be few consultants providing large quantities of data each. This would also decrease our total quantity of data, but others have been successful using methods similar to those

		beta				
	CER/WER	1	3	5	7	9
	0.25	77.3/100.0	74.9/100.0	**73.8**/100.0	74.5/100.0	78.2/100.0
alpha	0.5	80.7/100.0	77.7/100.0	75.2/100.0	74.3/100.0	75.0/100.0
	0.75	84.1/**98.6**	80.7/100.0	77.8/100.0	75.7/100.0	74.8/100.0

Table 3: The impact of tuning the language model parameters on Character and Word Error Rates for the baseline model.

		beta				
	CER/WER	1	3	5	7	9
	0.25	76.6/100.0	73.2/100.0	**72.1**/100.0	74.0/100.0	81.8/100.0
alpha	0.5	81.0/99.4	77.0/100.0	74.0/100.0	73.3/100.0	75.8/100.0
	0.75	85.2/**98.1**	80.1/100.0	77.2/100.0	74.8/100.0	74.7/100.0

Table 4: The impact of tuning the language model parameters on Character and Word Error Rates for the transfer learning model.

we outline above on smaller datasets (Meyer, 2019; Jimerson et al., 2018; Panaite et al., 2019; Yu et al., 2019).

The results in Table 2 show that the language model needs refinement, as it currently hinders rather than helps the performance of the system. The initial lm was trained on the training data from our corpus, and performed even worse than the current one. The current lm is assembled from a mix of domains from several time periods, which may be one explanation for its poor performance. However, tables 3 and 4 show that tuning the language model parameters is still important, and also indicate good parameters for training the neural network, as the language model is used for validation on the dev set.

8 Possible ELAN integration

Although the accuracy is at the moment rather low, it's worth considering how speech recognition technology could in principle be integrated into language documentation work. Previous work of (Gerstenberger et al., 2017a) presents a very effective approach to integrate a morphological analyser into ELAN through an external Python script, and there is no reason why speech recognition could not be implemented in similar fashion. The task may be computationally more complex, but if the speech recognition system is trained on individual utterances, it should always be possible to send such utterances as input to the system, and to predict their transcriptions.

From this point of view the most straightforward way to use speech recognition in this context could be to manually segment the ELAN file, as one normally does in manual workflows, and predict the transcription on each of those segments individually. In this paper we have only focused on the problem of speech recognition itself, but actually executing speech recognition on a new audio file involves segmentation and speaker diarization, both of which are complex and, to some degree, unsolved problems.

9 Conclusion & Further Work

The most central upcoming task is to repeat the experiment with other speech recognition systems that are currently available. Other potential lines of research would be to repeat this experiment with comparable datasets on other languages, in order to see whether the challenges reported in this paper are more connected to features of Komi dataset, or if they relate more to DeepSpeech infrastructure.

Meanwhile, there are also several things we can do towards improving the results on Komi. As several projects did report successful experiments when training on data that contains only an individual speaker, it seems logical to select only those speakers who contribute most to our corpus in the future, and retrain the system individually on that data. Similarly, simplifying the set of speakers such as male or female speakers only may have a similar effect.

Acknowledgments

Niko Partanen and Michael Rießler collaborate within the project Language Documentation meets Language Technology: The Next Step in the Description of Komi, funded by the Kone Foundation, Finland.

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation.

Rogier Blokland, Marina Fedina, Niko Partanen, and Michael Rießler. 2019. Spoken komi corpus. the language bank of finland version 1.0.0.

Ben Foley, Joshua T Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, et al. 2018. Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). In *SLTU*, pages 205–209.

Ben Foley, Alina Rakhi, Nicholas Lambourne, Nicholas Buckeridge, and Janet Wiles. 2019. Elpis, an accessible speech-to-text tool. *Proc. Interspeech 2019*, pages 4624–4625.

Fu-Lab. 2019. Корпус коми языка.

Ciprian Gerstenberger, Niko Partanen, and Michael Rießler. 2017a. Instant annotations in elan corpora of spoken and written komi, an endangered language of the barents sea region. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 57–66.

Ciprian Gerstenberger, Niko Partanen, and Michael Rießler. 2017b. Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 57–66, Honolulu. Association for Computational Linguistics.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition.

Anu-Reet Hausenberg. Komi. In Daniel Abondolo, editor, *The Uralic languages*, pages 305–326. Routledge.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197. Association for Computational Linguistics.

OO Iakushkin, GA Fedoseev, AS Shaleva, AB Degtyarev, and OS Sedova. 2018. Russian-language speech recognition system based on deepspeech.

Robbie Jimerson, Kruthika Simha, Raymond W Ptucha, and Emily Prudhommeaux. 2018. Improving ASR output for endangered language documentation. In *SLTU*, pages 187–191.

Marja Leinonen. 2002. Influence of Russian on the syntax of Komi. 57:195–358.

Marja Leinonen. 2006. The russification of Komi. Number 27 in Slavica Helsingiensia, pages 234–245. Helsinki University Press.

Marja Leinonen. 2009. Russian influence on the Ižma Komi dialect. *International Journal of Bilingualism*, 13(2):309–329.

Josh Meyer. 2019. Multi-task and transfer learning in low-resource speech recognition.

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit.

Marilena Panaite, Stefan Ruseti, Mihai Dascalu, and Stefan Trausan-Matu. 2019. Towards a Deep Speech model for Romanian language. In *2019 22nd International Conference on Control Systems and Computer Science (CSCS)*, pages 416–419. IEEE.

Jillur Rahman Saurav, Shakhawat Amin, Shafkat Kibria, and M Shahidur Rahman. 2018. Bangla speech recognition for voice search. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4. IEEE.

Chongchong Yu, Yunbing Chen, Yueqiao Li, Meng Kang, Shixuan Xu, and Xueer Liu. 2019. Cross-language end-to-end speech recognition research based on transfer learning for the low-resource Tujia language. *Symmetry*, 11(2):179.

ЛМ Безносикова, ЕА Айбабина, НК Забоева, and РИ Коснырева. 2012. Коми сёрнисикас кывчукӧр. Словарь диалектов коми языка: в 2-х томах/ИЯЛИ Коми НЦ УрО РАН; под ред. ЛМ Безносиковой.

Йӧлгинь Цыпанов. 2009. Перым кывъяслӧн талунъя серпас. *Suomalais-Ugrilaisen Seuran Toimituksia = Mémoires de la Société Finno-Ougrienne*, 258:191–206.