Smit, Peter; Leinonen, Juho; Jokinen, Kristiina; Kurimo, Mikko

Automatic Speech Recognition for Northern Sámi with comparison to other Uralic Languages

Published in:
Proceedings of the Second International Workshop on Computational Linguistics for Uralic Languages

Published: 20/01/2016

Document Version
Peer reviewed version

Please cite the original version:
Smit, P., Leinonen, J., Jokinen, K., & Kurimo, M. (2016). Automatic Speech Recognition for Northern Sámi with comparison to other Uralic Languages. In Proceedings of the Second International Workshop on Computational Linguistics for Uralic Languages (pp. 80-91). [9] Szeged, Hungary: University of Szeged.

Automatic Speech Recognition for Northern Sámi with comparison to other Uralic Languages

Peter Smit[1]
peter.smit@aalto.fi

Juho Leinonen[1]
juho.leinonen@aalto.fi

Kristiina Jokinen[2]
kristiina.jokinen@helsinki.fi

Mikko Kurimo[1]
mikko.kurimo@aalto.fi

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Institute of Behavioural Sciences, University of Helsinki, Finland

December 30, 2015

Abstract

Speech technology applications for major languages are becoming widely available, but for many other languages there is no commercial interest in developing speech technology. As the lack of technology and applications will threaten the existence of these languages, it is important to study how to create speech recognizers with minimal effort and low resources.

As a test case, we have developed a Large Vocabulary Continuous Speech Recognizer for Northern Sámi, an Finno-Ugric language that has little resources for speech technology available. Using only limited audio data, 2.5 hours, and the Northern Sámi Wikipedia for the language model we achieved 7.6% Letter Error Rate (LER). With a language model based on a higher quality language corpus we achieved 4.2% LER. To put this in perspective we also trained systems in other, better-resourced, Finno-Ugric languages (Finnish and Estonian) with the same amount of data and compared those to state-of-the-art systems in those languages.

1 Introduction

The field of speech recognition is maturing, as companies start to actively use and sell products that utilize Large Vocabulary Continuous Speech Recognition (LVSCR). Especially the creators of operating systems for mobile devices incorporate methods into their products to operate devices using voice.

These commercial applications however, are only focusing on small fraction of the languages in the world. Other languages do not have the required data and expertise readily available, and are therefore left out from these systems as it would not be commercially viable to create these applications. Especially minority languages and languages from developing countries receive only minor academic and commercial interest for the development of LVCSR systems. [1]

One for these under-resourced languages is Northern Sámi, the largest of the nine Sámi languages with approximately 25,000 speakers. It belongs to the Uralic language family. [2]

Like other languages in the Finno-Ugric branch of the Uralic language family, e.g. Finnish and Estonian, it is a highly morphological language that uses independent suffixes extensively. This poses challenges for speech technology applications as the number of inflections, derivations and compoundings cause the size of the vocabulary to be enormous, especially compared to the languages in the Indo-European family [3]. A large vocabulary especially causes problems in the estimation of language models, which can not produce any words beyond those seen in the the training data.

Northern Sámi is an under-resourced language, as there are little corpora of spoken and written language available, and financial resources to collect these data are limited. Even though there is active linguistic research on Northern Sámi, there are limitations to the expert resources available for speech recognition, such as pronunciation dictionaries.

To combat the challenges of building an LVCSR system for an under-resourced language we have employed several techniques. First we used 'found data' for building the acoustic and language models. For the acoustic model we bootstrapped from a better resourced related language (Finnish). For the language model we increased the coverage of the model by employing sub-word units (morphs) instead of words. Similar techniques have been used in [1, 4] but here we wanted to evaluate their applicability to uralic languages, in particular. This work is an extension of [5].

Because there are no state-of-the-art LVCSR references for Northern Sámi, we simulated the potential of larger resources by studying also two better resourced Uralic languages. First we produced systems for Finnish and Estonian using the corresponding data as we had for Northern Sámi. Then we compared these systems to similar systems that we produced using larger data and, finally, to the state-of-the art systems for these languages. These results helped us to estimate the gains for collecting more Northern Sámi data.

1.1 WikiTalk and DigiSami

Another motivation to build a recognizer for Northern Sámi is to utilize it as part of a spoken dialogue system in the WikiTalk application [6]. This is one part of the DigiSami project which is a research project at University of Helsinki aiming to support content generation of less resourced languages with the help of language technology. Currently, the main dangers to Sámi language are the disappearance of the traditional lifestyle and work of Sámi culture, and emigration of Sámi people away from their old living areas. However, there are also studies and discussion on using new technologies to revitalize languages [7]. In [8], revitalization for the Northern Sámi language is described using spoken language data collection in interactive setting for the WikiTalk application. In WikiTalk, the idea is to have users (children or adults) find out more about subjects that interest them by discussing with the humanoid robot Nao. They can ask for more information on the subject and then Nao will read them the related Wikipedia article [9]. Described in this paper is the first step to building this end-to-end system.

2 ASR for under-resourced languages

The majority of the state-of-the art methods in Large Vocabulary Speech Recognition require large amounts of data and expertise.

Firstly, a great number of high quality spoken utterances have to be collected and correctly transcribed. For a Speaker-Independent (SI) system, i.e. a system that can recognize anyone who speaks the target language, utterances from many different persons are needed. For a Speaker-Dependent

(SD) system, i.e. a system that can only recognize the voice of the person who provided the training data, only a few hours of transcribed speech are required.

The second required dataset is a large corpus of written text, preferably in the same style and domain as what should be recognized by the system. Te corpus is used to train the language model and it should contain all common words in their expected contexts.

Lastly, a speech recognizer needs a pronunciation dictionary; i.e. a list of all possible words with all their possible phonetic transcriptions. The phonemes also need to be grouped according to different phonetic properties, so that their probability distributions can be shared in the training of the acoustic model.

For under-resourced languages, as the name suggests, none of the above data is readily available, but alternative solutions have to be developed. An easy alternative to a large corpus of transcribed audio data is to collect audio books. Although the quality of the speech varies, projects such as Librivox have freely available audio books in many languages which can be used for this purpose. Using a temporary acoustic model and simple text processing techniques these audio books can be automatically segmented into sentence-long utterances that are suitable for training a minimal speaker-dependent model.

Language data is also freely available on-line, and e.g. Web-scraping can give a rudimentary dataset for training a language model [10]. Also sites like Wikipedia have often big collections of easily available text. However, the quality and usability of such data varies, and many of the sources that can be 'found' on-line suffer from the problem that their style and topic are non-standard and do not necessarily match written nor spoken language conventions. Moreover, on-line texts often contain foreign language segments, symbols or abbreviations which decrease their usability for building language models.

One of the main resource consuming tasks is the preparation of a pronunciation dictionary, which normally requires extensive manual work and linguistic knowledge. One solution to build the pronunciation dictionary quickly is to model the graphemes (letters) of the words directly, instead of using the actual phone they represent [11]. In languages such as English this does not, of course, give very good results since graphemes can have very different realizations. Consider for example the words 'tough' and 'dough' that resemble each other in writing, but are pronounced in a completely different way. In the Uralic languages studied here however, a grapheme-to-sound system works reasonably well since, in general, every grapheme is realized as a single distinct sound.

Lastly, the phonetic grouping or 'phoneme question set' is a small dataset that requires linquistic expertise. Although there are algorithms available that can replace this set altogether [12], it is often undesirable as it makes the system less effective. It is also possible to modify the phoneme set of a closely related language, and such small modifications to approximate the target language do not necessarily require so much expert effort.

Even though the above simplified solutions can replace all the expensive data needs, they will inevitably limit the performance of the speech recognizer. Adding more and domain related training data as well as developer expertise will naturally improve the system performance significantly. However, the low-resource systems can already serve some basic language technology needs. The largest limiting factor for these systems is that a real SI system requires training data from more than a hundred speakers.

3 Acoustic modeling

The Acoustic Modeling part of the speech recognizer was done with a standard Hidden Markov model with Gaussian mixture models as emission distribution (HMM-GMM). Mel Frequency Cepstral Coefficients were used as input features. [13]

The audio data is prepared by splitting the audio files (originally chapter length or similar) into sentence utterances. This is done by doing Baum-Welch forced alignment with a temporary speech recognition model. The temporary speech recognition model was created by taking a well trained Finnish model and mapping the Finnish phonemes to the one of the target language. In later iterations the best speech recognition model of the language was used to do the forced alignment again, resulting in a perfect split of training utterances.

The HMM-GMM model is trained using multiple iterations of Baum-Welch maximum likelihood estimation. To manage the model complexity Gaussians were shared between different HMM-states using decision tree clustering. The modeling unit of the acoustic model is a tri-state tri-phone, which means that all the phonemes with a different preceding and succeeding phoneme are modeled as separate units, as are the beginning, middle and end of each tri-phone.

In Section 7 the number of Gaussians for different models are reported.

4 Language modeling

A language model is an important part of any speech recognition system. Even though theoretically a good acoustic model with a lexicon could be enough to recognize words, a model that takes the word context is essential. For languages which have many homophones, i.e. words with the same pronunciation but different meaning, it is also essential to have a language model, so as to pick the right word meaning given a pronunciation in the context.

A language model predicts words based on their sentence context. For synthetic languages like Finnish and all the Uralic languages, the main issue with word-based language modeling is that a huge lexicon is needed in order to decrease the out-of-vocabulary (OOV) rate to a manageable percentage. Since the OOV-rate is the minimum WER possible, an OOV-rate much less than 10% is necessary. For an English speech recognizer, a vocabulary size of 20 000 word may provide an OOV-rate of 2.4-2.7%, while with a vocabulary of 40 000 words, an OOV-rate less than one percent is achieved [14]. In contrast, a Finnish recognizer needs a 410 000 word vocabulary to have an OOV-rate of 4.0-7.3% [15].

An interesting alternative for a word-based language model is to use a sub-word language model. A sub-word model builds words out of a smaller set of word fragments. The word fragments are particularly effective in agglutinative languages or languages with a lot of compound words. When the words are built from smaller units, also the OOV words can be modelled by using the probabilities of sub-word unit combinations learned from the training corpus. If the word fragments are chosen appropriately, the OOV-rate can become close to zero, even for smaller language data corpora.

4.1 Morfessor

Morfessor is a machine learning tool that uses a statistical model to split words into smaller fragments, which can be used for language modeling [16]. This resembles closely the splitting of words into their smallest informational units, morphemes.

Morfessor has three components; the model, the cost function, and the training and decoding algorithms. The model contains the lexicon, i.e. the properties of the morphs, the written form of the morph itself and its frequency, as well as the grammar, which contains information of how the morphs can be combined into words. The Morfessor cost function is derived from a MAP estimation with the goal of finding the optimal parameters $\boldsymbol{\theta}$ given the observed training data $\boldsymbol{D}_W$:

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\theta} P(\boldsymbol{\theta}|\boldsymbol{D}_W) = \arg\max_{\theta} P(\boldsymbol{\theta})P(\boldsymbol{D}_W|\boldsymbol{\theta}). \tag{1}$$

The cost function to be minimized is the negative logarithm of the product $P(\boldsymbol{\theta})P(\boldsymbol{D}_W|\boldsymbol{\theta})$

$$L(\boldsymbol{\theta}, \boldsymbol{D}_W) = -\log P(\boldsymbol{\theta}) - \log P(\boldsymbol{D}_W|\boldsymbol{\theta}). \tag{2}$$

The purpose of this is to generate a small set of morphs that represents the words in the training corpus compactly. If only letters were used as morphs the set of would be small but representing the corpus with individual letters would be cumbersome. In contrast using whole words as morphs would result in a large set of morphs so the optimal solution is somewhere in between. However, individual letters are added to the morph set so even previously unseen words can always be segmented.

A greedy search algorithm is used to find the optimal segmentation of morphs for the training data. When the best model is found, it is used to segment the language model training corpus with the Viterbi algorithm. This result can be used to generate n-gram models with morphs as LM units.

4.2 n-gram modeling

n-gram models predict the output of the next word or sub-word given the $n-1$ previous words or sub-words. They are normally created by counting all occurrences of the word and sub-word sequences. To prevent the model from being too big and too much tailored to the training data (overfitting), pruning is applied. Also, some of the probability mass is reserved for unseen contexts, for example with the Kneser-Ney smoothing technique[17].

When n-gram models are build for words, the order the model, i.e the value of n, is typically between three and five. If the order is high, the models get too big, and they do not contain enough necessary information. With the sub-word models, however, the contexts can be much deeper, as there are less types in the vocabulary and the context counts are more sparse. Also intuitively, to cover the same context, the order of a sub-word model must be higher than the order or the word model. Standard tools for n-gram modelling have problems with correctly smoothing and growing high-order n-gram language models. VariKN [18] is a specific algorithm and tool to solve this problem and it was used in this paper for building high-order sub-word n-gram models.

5 Experiment setup

The experiments were carried out using our open source speech recognition toolkit called AaltoASR[1] [13][19]. It uses context-dependent tri-phones with diagonal Gaussian mixture models (GMM) as emission distributions and the speech features itself are Mel-Frequency Cepstral Coefficients (MFCCs).

[1]Open source, available from `https://github.com/aalto-speech/AaltoASR`

Both words and sub-word units were used for language modeling. The sub-word unit models were created with Morfessor 2.0[2], an implementation of the Morfessor Baseline algorithm[20].

Variable length n-grams used for language modeling were generated by both SRILM[3] [21] and VariKN[4] [18, 22]. The decoder of AaltoASR is a time-synchronous one-pass token passing decoder where the beam search is complemented by a language model look-ahead [23].

6 Northern Sámi ASR evaluation

The audio data used for the Northern Sámi recognizer came from the UIT-SME-TTS corpus[5]. There are data for two speakers, one male and one female. The male audio data was 4.7 hours and the female data 3.3 hours. Separate data is needed for development and evaluation, and we used 75% for training. This makes 3.5 and 2.5 hours of training data for the male and female voice, respectively.

The initial recognition model was created by using a Finnish model. With this model, the audio data was split into sentences and trained with the procedure described in Section 3. This resulted in two speaker dependent systems, one for the male and one for the female speaker (resp. SM1 and SF1). These models are Speaker-Dependent models as there is data only from the two speakers available.

For language model, we evaluated both word and morph n-gram models. In addition to the training sentences, we also used the Northern Sámi Wikipedia dump (Train+Wiki).

The results for basic recognition are shown in Table 1. Besides the standard Word Error Rate (WER), also the Letter Error Rate (LER) is reported. LER is common for speech recognition experiments on languages which are morphological complex such as Northern Sámi, Finnish and Estonian.

		Speaker SF1			Speaker SM1		
Unit	**Toolkit**	**5-gram**	**7-gram**	**9-gram**	**5-gram**	**7-gram**	**9-gram**
words	SRILM	52.9 / 12.7	52.9 / 12.7	52.9 / 12.7	48.6 / 11.1	48.7 / 11.1	48.7 / 11.1
morphs	SRILM	40.0 / 9.0	39.9 / 9.3	39.1 / 9.1	37.6 / 8.5	36.8 / 8.4	37.3 / 8.5
morphs	VariKN	38.4 / 8.6	38.5 / 8.7	37.6 / 8.7	35.4 / 8.1	33.7 / 7.6	34.1 / 7.9

Table 1: ASR recognition results for the Northern Sámi SD recognizers. Word Error Rate / Letter Error Rate reported.

We first observe that the SM1 recognizer is slightly better than the SF1 recognizer. However, this is most likely caused by the fact that there was more data available for the training of the acoustic model.

As expected, the morph based language models have much lower error rates than the word-based models. Looking at Table 2, we notice that the OOV-rate for word based models is rather high which causes the big difference in performance to sub-word models.

For word-based models there is no effect on using higher order n-grams. This can be seen in Figure 1 which shows WER for different n-gram models with the SM1 system. In this comparsion we used the Big Northern Sámi language model which is trained from approximately 12 million word tokens of data from 'Den samiske textbanken'. There is no change in performance after the 3rd order n-grams

[2]Open source, available from `http://www.cis.hut.fi/projects/morpho/`

[3]Open source, available from `http://www.speech.sri.com/projects/srilm/`

[4]Open source, available from `https://github.com/vsiivola/variKN`

[5]Provided by the University of Tromsø

	Word	Morph
Female	22%	0%
Male	20%	0%

Table 2: Out-of-vocubulary percentages for the Female and Male testsets.

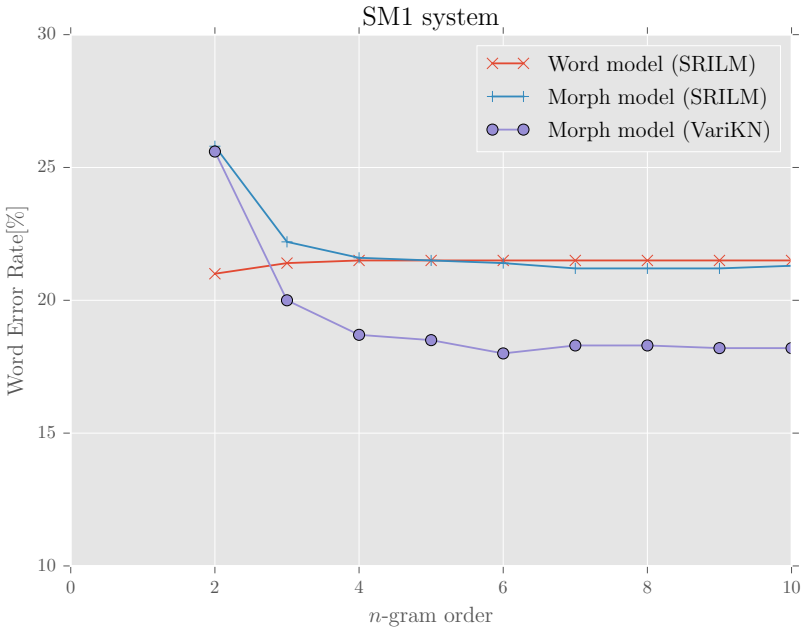

Figure 1: Word error rates for the SM1 system with the Big language model.

for the the word-based model, whereas for the VariKN morph-based models there are clear effects when using higher order models.

The best Word Error Rate on the Big language model for the SM1 system is 18.2%, the best Letter Error Rate 4.2%.

7 Comparison of low-resource systems for multiple languages

To compare the results of the Northern Sámi recognizer with recognizers in different languages we first train Speaker Dependent models for both Finnish and Estonian audio books. The available audio datasets are described in Table 3 and the available text corpora in Table 4.

Even though all datasets are audio books, there are a number of differences. EF1, EM1 and FM1 were encoded with the mp3-codec, while for the FF1, SM1 and SF1 audio books original high quality uncompressed audio files were available. The speaking style was generally the same, with a prosody typical to story telling. An exception to this was the FF1 book, an audio book created for blind persons, which has been read in a very monotone voice with little prosodic variation. This makes the book also understandable when played at higher speeds.

[6]Provided by YLE. Can be listened on `http://areena.yle.fi/1-1301621`

	Language	Gender	Title	Amount
EF1	Estonian	Female	Nils Holgerssoni imeline teekond läbi Rootsi	16 hours
EM1	Estonian	Male	Würst Gabriel ehk Pirita kloostri wiimsed päewad	6 hours
FF1	Finnish	Female	Syntymättömien sukupolvien Eurooppa	12 hours
FM1	Finnish	Male	Seitsemän veljestä[6]	13 hours
SF1	Northern Sámi	Female	UIT-SME-TTSF	3.3 hours
SM1	Northern Sámi	Male	UIT-SME-TTSM	4.6 hours

Table 3: Audio data for the trained speaker dependent systems.

Language	Source	#sentences	#word tokens	#word types
Estonian	Wikipedia	895k	10M	778k
Estonian	newspaper+web+broadcast [24]	19M	229M	3.8M
Finnish	Wikipedia	2.2M	22M	1.5M
Finnish	Kielipankki	13M	143M	4.1M
Northern Sámi	Wikipedia	10k	88k	20k
Northern Sámi	Den samiske tekstbanken	990k	12M	475k

Table 4: Language modeling data for the trained speaker dependent systems.

The experiments in Section 6 confirmed the hypothesis that morph-based n-gram models trained with the VariKN toolkit give the best performance, hence only this combination will be used.

To compare the systems for different languages fairly, we artificially reduce the amount of audio and text data to match that of our smallest system. We only use 2.5 hours of audio data and a random 10.000 sentences of the Wikipedia data set for each language. The systems are trained with a 10-gram VariKN sub-word language model. The statistics in Table 5 show that the datasets have equal number of sentences, but not equal number of word types or tokens. This is most likely due to the Northern Sámi Wikipedia having more stub articles that contain short sentences with similar words.

The Train+Wiki language models are trained from the combination of the recognizer's training sentences and the small Wikipedia dataset as described in Table 5. The Big language models are trained from the higher quality text sources, which are described in Table 4.

The results of the comparable systems with the Train+Wiki dataset are shown in Table 6. The word error rates are close to each other, confirming that the systems are comparable. One exception is the FF1 system, which performs much better. This better result is most likely a combination of the speaking style, which had little variation, and a better match between the text of the language model and the test data.

We also tested the models with the same amount of acoustic data and their respective Big language models. The improvements are significant with the best improvement being 64% relative improvement in WER for the FF1 system. This indicates the importance of the availability of high quality language model data for the performance of a Uralic speech recognition system. The amount of data

Language	#sentences	#word tokens	#word types
Estonian	10k	108k	41k
Finnish	10k	103k	43k
Northern Sámi	10k	88k	20k

Table 5: Reduced subsets of wikipedia data for use in the Train+Wiki language model.

Language	Voice	Train+Wiki WER	Train+Wiki LER	Big WER	Big LER
Estonian	EF1	39.6	15.8	25.0	11.4
Estonian	EM1	39.2	13.3	25.5	9.6
Finnish	FF1	25.2	4.1	8.9	2.1
Finnish	FM1	35.8	7.7	24.9	5.6
Northern Sámi	SF1	37.5	8.5	23.7	5.5
Northern Sámi	SM1	39.5	9.4	20.9	4.9

Table 6: Word Error Rates for using 2.5 hours of training data and either the Train+Wiki or Big language models. All language models were 10-gram VariKN sub-word models.

Language	Voice	#hours	#Gaussians	2.5 hours WER	2.5 hours LER	All data WER	All data LER
Estonian	EF1	8	31.5k	25.0	11.4	18.8	8.3
Estonian	EM1	4.5	12.6k	25.5	9.6	23.2	8.4
Finnish	FF1	9	26k	8.9	2.1	8.1	1.9
Finnish	FM1	10	28k	24.9	5.6	19.8	3.7
Northern Sámi	SF1	2.5	7.7k	23.7	5.5	23.7	5.5
Northern Sámi	SM1	3.5	9.6k	20.9	4.9	18.1	4.2

Table 7: Speech recognizer results for the full audio books with the Big language model.

however is less important, as the Big language model for Northern Sámi gives a similar improvement as the Big language models for the other systems, even though the amount of data in the Big language model for Northern Sámi is lower than the amount of data in the Train+Wiki systems for Estonian and Finnish.

To see the effect of using more acoustic data, we also trained all systems on their full acoustic datasets and evaluated them with the Big language model. While the 2.5 hour data systems were all modeled with appr. 7,500 Gaussians, the bigger models have proportionally more Gaussians.

The results are shown in Table 7. There are a couple of surprising results. For the FF1 system, there is a small improvement on the already very good result. On the other hand, the SM1 system already improves with 13% relative WER with only an hour of added data. In general, there is a clear pattern that more acoustic data improves the model, except if the data has so little variation that an optimum is already reached earlier.

7.1 State-of-the-art recognizers

The experiments in the previous sections show that results on Finnish and Estonian systems are comparable with Northern Sámi systems if the same amount of data is provided. This allows us to look to the state-of-the-art recognition systems for Finnish and Estonian systems and project how well a Northern Sámi system would perform if the same amount of data would be collected.

Table 8 shows the reported error rates for different systems. The most important difference with the systems discussed in the previous sections is that these are Speaker Independent recognizers, which are tested with different speakers than those present in the training data. Also the quality and type of speech are different.

Of these state-of-the-art results, the results on the Finnish Speecon set and the Finnish telephone

Language	Description	WER	LER	Source
Estonian	Broadcast conversations	17.9%		[25]
Estonian	Oral presentations	26.3%		[25]
Finnish	Speecon testset		2.9%	[26]
Estonian	Telephone speech	33.1%	11.9%	[13]
Finnish	Telephone speech	21.6%	6.8%	[13]

Table 8: State of the art results for Finnish and Estonian Speaker Independent ASR.

speech are most impressive. Even though there is much more speaker variability, the result on the Speecon testset is close to the result of the FF1 SD recognizer. This is done using speaker adaptive training and discriminative training techniques.

The telephone speech results are focused on lower quality speech data. Again the results seem better for the SD systems in the previous section, but the variability in speakers, the speech quality and the language content of the utterances are much more complex.

Given that the Speaker Dependent systems all performed with similar accuracy, we expect that tasks of similar difficulty would perform as well for Northern Sámi as they would for Finnish or Estonian, given that the data would be available.

8 Conclusions

Using a number of techniques, most notably sub-word language models and grapheme-to-sound acoustic modeling, we have overcome challenges caused by a small amount of data available for developing speech recognizer systems for under-resourced languages. We have demonstrated the feasibility of this approach by training Speaker Dependent speech recognizer systems for the Northern Sámi language, an under-resourced Finno-Ugric language, and achieved a letter error rate of only 4.2%.

In order to put the result in perspective and validate the techniques, we also trained systems for Finnish and Estonian using artificially limited datasets. These experiments show that the Northern Sámi recognizer gives comparable results to the Finnish and Estonian recognizers and can effectively use similar techniques such as sub-word language models.

In future work we plan to use cross-lingual techniques to build Speaker Independent systems for Northern Sámi, even though acoustic datasets with enough different speakers might not be available, or only available without transcriptions.

All scripts used in this paper are published as open-source under the Modified BSD license[7].

9 Acknowledgements

We thank the University of Tromsø for the access to their Northern Sámi datasets. This research has been supported by the Academy of Finland under the Finnish Centre of Excellence Program 2012–2017 (grant n°251170, COIN) , and through the project Fenno-Ugric Digital Citizens (grant n°270082). We acknowledge the computational resources provided by the Aalto Science-IT project.

[7]Available from `https://github.com/phsmit/iwclul2016-scripts`

References

[1] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100, 2014.

[2] M. Paul Lewis, Gary F. Simons, and Charles D. (eds.) Fennig. Ethnologue: Languages of the world, eighteenth edition. Online version: http://www.ethnologue.com., 2015.

[3] Fred Karlsson. *Suomen kielen äänne- ja muotorakenne*. WSOY, Helsinki, 1982.

[4] Viet-Bac Le and L. Besacier. Automatic speech recognition for under-resourced languages: Application to Vietnamese language. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(8):1471–1482, Nov 2009.

[5] Juho Leinonen. Automatic speech recognition for human-robot interaction using an under-resourced language. Master's thesis, Aalto University School of Electrical Engineering, Espoo, 2015.

[6] Graham Wilcock and Kristiina Jokinen. Wikitalk human-robot interactions. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI)*, pages 73–74. ACM, 2013.

[7] Patrick Eisenlohr. Language revitalization and new technologies: Cultures of electronic mediation and the refiguring of communities. *Annual Review of Anthropology*, pages 21–45, 2004.

[8] Kristiina Jokinen. Open-domain interaction and online content in the Sami language. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2014)*, 2014.

[9] Kristiina Jokinen and Graham Wilcock. Multimodal open-domain conversations with the Nao robot. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 213–224. Springer, 2014.

[10] Kevin P Scannell. The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15, 2007.

[11] Stephan Kanthak and Hermann Ney. Multilingual acoustic modeling using graphemes. In *INTERSPEECH*, pages 1145–1148, 2003.

[12] Sakhia Darjaa, Milos Cernak, Marián Trnka, Milan Rusko, and Róbert Sabo. Effective triphone mapping for acoustic modeling in speech recognition. In *INTERSPEECH*, pages 1717–1720, 2011.

[13] Teemu Hirsimäki, Janne Pylkkönen, and Mikko Kurimo. Importance of high-order n-gram models in morph-based speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(4):724–732, 2009.

[14] Phil C Woodland, CJ Leggetter, JJ Odell, V Valtchev, and SJ Young. The 1994 HTK large vocabulary speech recognition system. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 73–76. IEEE, 1995.

[15] Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language*, 20(4):515–541, 2006.

[16] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):3, 2007.

[17] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics, 1996.

[18] Vesa Siivola, Teemu Hirsimaki, and Sami Virpioja. On growing and pruning Kneser–Ney smoothed-gram models. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(5):1617–1624, 2007.

[19] Janne Pylkkönen. An efficient one-pass decoder for Finnish large vocabulary continuous speech recognition. In *Proceedings of The 2nd Baltic Conference on Human Language Technologies*, pages 167–172, 2005.

[20] Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. Morfessor 2.0: Python implementation and extensions for Morfessor baseline. Technical report, 2013.

[21] Andreas Stolcke et al. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*, 2002.

[22] Vesa Siivola, Mathias Creutz, and Mikko Kurimo. Morfessor and variKN machine learning tools for speech and language technology. In *INTERSPEECH*, pages 1549–1552, 2007.

[23] Stefan Ortmanns, Andreas Eiden, Hermann Ney, and Norbert Coenen. Look-ahead techniques for fast beam search. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 3, pages 1783–1786. IEEE, 1997.

[24] Mikko Kurimo, Seppo Enarvi, Ottokar Tilk, Matti Varjokallio, André Mansikkaniemi, and Tanel Alumäe. Modeling under-resourced languages for speech recognition. *Language Resources and Evaluation*, in review.

[25] Tanel Alumäe. Recent improvements in Estonian LVCSR. In *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[26] Janne Pylkkonen and Mikko Kurimo. Analysis of extended Baum-Welch and constrained optimization for discriminative training of hmms. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(9):2409–2419, Nov 2012.