Leinonen, Juho; Smit, Peter; Virpioja, Sami; Kurimo, Mikko

New Baseline in Automatic Speech Recognition for Northern Sámi

Published in:
Fourth International Workshop on Computational Linguistics for Uralic Languages

DOI:
10.18653/v1/W18-0208

Published: 01/01/2017

Document Version
Peer reviewed version

Please cite the original version:
Leinonen, J., Smit, P., Virpioja, S., & Kurimo, M. (2017). New Baseline in Automatic Speech Recognition for Northern Sámi. In Fourth International Workshop on Computational Linguistics for Uralic Languages (pp. 89-99). ACL. https://doi.org/10.18653/v1/W18-0208

New Baseline in Automatic Speech Recognition for Northern Sámi

Juho Leinonen
Aalto University
juho.leinonen@aalto.fi

Peter Smit
Aalto University
peter.smit@aalto.fi

Sami Virpioja
Utopia Analytics
Aalto University
sami.virpioja@aalto.fi

Mikko Kurimo
Aalto University
mikko.kurimo@aalto.fi

Abstract

Automatic speech recognition has gone through many changes in recent years. Advances both in computer hardware and machine learning have made it possible to develop systems far more capable and complex than the previous state-of-the-art. However, almost all of these improvements have been tested in major well-resourced languages. In this paper, we show that these techniques are capable of yielding improvements even in a small data scenario. We experiment with different deep neural network architectures for acoustic modeling for Northern Sámi and report up to 50% relative error rate reductions. We also run experiments to compare the performance of subwords as language modeling units in Northern Sámi.

Tiivistelmä

Automaattinen puheentunnistus on kehittynyt viime vuosina merkittävästi. Uudet innovaatiot sekä laitteistossa että koneoppimisessa ovat mahdollistaneet entistä paljon tehokkaammat ja monimutkaisemmat järjestelmät. Suurin osa näistä parannuksista on kuitenkin testattu vain valtakielillä, joiden kehittämiseen on tarjolla runsaasti aineistoja. Tässä paperissa näytämme että nämä tekniikat tuottavat parannuksia myös kielillä, joista aineistoa on vähän. Kokeilemme ja vertailemme erilaisia syviä neuroverkkoja pohjoissaamen akustisina malleina ja onnistumme vähentämään tunnistusvirheitä jopa 50%:lla. Tutkimme myös tapoja pilkkoa sanoja pienempiin osiin pohjoissaamen kielimalleissa.

1 Introduction

The field of automatic speech recognition (ASR) has advanced rapidly in the last couple of years, in large part thanks to deep neural networks (DNNs). For decades there

has been active research trying to replace Gaussian mixture models (GMM) with various neural network configurations. Yet, only after 2010 the full power of neural networks started to be noticed when multiple groups started reporting huge improvements in their implementations (Hinton et al., 2012). At the same time, the computational power of modern graphics processing units (GPU) has made it feasible to utilize very large DNNs with very large training data sets. For speech recognition, this has meant that the decades-old best practices are quickly being replaced by new and more powerful methods.

In this paper, we have documented our work to build a new baseline for Northern Sámi. Using DNNs for acoustic modeling has provided large improvements for well-resourced Uralic languages, but for under-resourced languages, the applicability has yet to be tested. For broadcast news data sets, the latest improvements for applying neural networks instead of GMM-based acoustic models have been in the range of 14% smaller relative word error rate (WER) for Finnish and 6% for Estonian (Smit et al., 2017b).

In languages with a rich morphological structure it is difficult to build statistical language models using words. If using n-gram word models, the vocabulary size becomes computationally challenging, and even worse, the growing lexicon decreases out-of-vocabulary (OOV) rate rather slowly. Furthermore, the lack of data for under-resourced languages makes building a large lexicon and n-gram difficult. For Finnish, Estonian, Arabic and Turkish it is common to use subword units such as morphs (Hirsimäki et al., 2006) or syllables (Choueiter et al., 2006) instead of words. In this work we follow this tradition and apply statistical morphs as subword units for Northern Sámi.

Because the pronunciation in Northern Sámi can be rather well covered by rules, a simple grapheme-to-phoneme conversion can be applied for our lexicon. This gives Northern Sámi and other such languages a significant advantage in ASR, since building a proper lexicon is one of the most arduous data preparation tasks for speech recognition.

We will use a popular open-source toolkit for speech recognition, Kaldi, and document the building of a speech recognizer. In addition to DNN-based acoustic modeling, we test new methods of subword modeling for morphologically rich languages, originally developed for Finnish. The main focus of the paper is to demonstrate these new techniques in building a new baseline for Northern Sámi for further research and comparison. We will compare our results to the previous Northern Sámi baseline results from Smit et al. (2016).

2 Methods

Our baseline system builds on the Northern Sámi recognizer by Smit et al. (2016), but with a few important changes. In acoustic modeling, we model triphones by hidden Markov models with Gaussian mixture model emission distributions (GMM-HMM) using mel frequency cepstral coefficients (MFCCs) as input features. The lexicon is based on subword units found by a data-driven method, and a long-context n-gram model is used for language modeling. However, while Smit et al. (2016) used the token-pass decoder of the AaltoASR toolkit (Pylkkönen, 2005; Hirsimäki et al., 2009),

our system is based on the Kaldi toolkit (Povey et al., 2011) that has a decoder based on weighted finite-state transducers (WFST). Kaldi has also implemented quite a few improvements to the standard GMM-HMM methodology. To further improve the speech recognition accuracy in Northern Sámi we test recent developments on creating subword lexicon for Kaldi and acoustic modeling based on DNNs.

2.1 WFST-based speech recognition

Kaldi is an open source toolkit for speech recognition developed since the year 2009 by researchers from many different universities, lead by the John Hopkins University and Brno University of Technology (Povey et al., 2011). It is based on the use of weighted finite-state transducers (WFST) complimenting the work by Mohri et al. (2008). The advantage of WFST-based recognizers is that once the search network has been constructed and optimized effectively by the WFST methods, the decoding is very fast and accurate. Moreover, Kaldi's GMM-HMMs are improved by subspace Gaussians, word-position-dependent phones and advanced silence models.

2.2 Subword lexicon FSTs and language models

The small amount of training data and the morphological complexity of Northern Sámi make it problematic to build language models (LM) using words as the basic units. We applied the data-driven Morfessor Baseline method (Creutz and Lagus, 2002, 2007) to segment the words into subword units. Because all words in the language can be composed from these subword units, this approach provides an unlimited vocabulary for ASR (Hirsimäki et al., 2006). While Morfessor was developed to find units of language that resemble the surface forms of linguistic morphemes, the current implementation includes a parameter for adjusting the level of segmentation that the method produces (Virpioja et al., 2013). The optimal level of segmentation for ASR varies between languages, but a wide range of lexicon seems to produce near-optimal results (Smit et al., 2017b). We did not experiment with this parameter.

Recently, Smit et al. (2017b) implemented effective subword modeling in the WFST-based ASR framework. It modifies the basic lexicon FST by introducing different models for all four different positions where a subword can appear (as prefix, infix, suffix, or complete word) and provides the appropriate word-position-dependent phones. In Figure 1 a normal word lexicon is shown where $words is replaced by a linear FST of all pronunciations in the lexicon. In Figure 2 the same basic structure is shown for a subword lexicon.

When the ASR system uses a subword lexicon, the subword units in the output need to be joined back to construct complete word forms. This can be accomplished in different ways; popular approaches are using a separate word boundary units (e.g. Hirsimäki et al., 2009) or using a special character to indicate that there is no word boundary directly preceding the subword (e.g. Arisoy et al., 2009; Tarján et al., 2014). Smit et al. (2017b) experimented on different styles of subword markings and the conclusion was that the optimal boundary marking style might depend on the language. Other work by the same authors (Smit et al., 2017a) supports this hypothesis. Therefore, in this work, we also experiment on different boundary marking styles to select the one that fits best for Northern Sámi. In Table 1 the four possible styles of marking

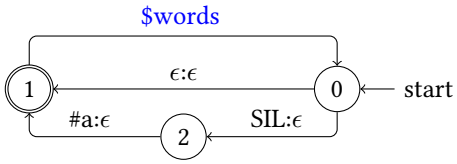

Figure 1: Prototype Lexicon FST for word-based lexicon. On each vertice in this graph is shown an input and output symbol. For example 'SIL:ϵ' indicates a SIL phone as input and a skip-token (ϵ) as output. The symbol #a is a disambiguation symbol which is required in Kaldi to make the FST determinizable. $words is a placeholder that is supposed to be replaced by a linear FST that maps all words to their appropriate word-position dependent phoneneme sequences.

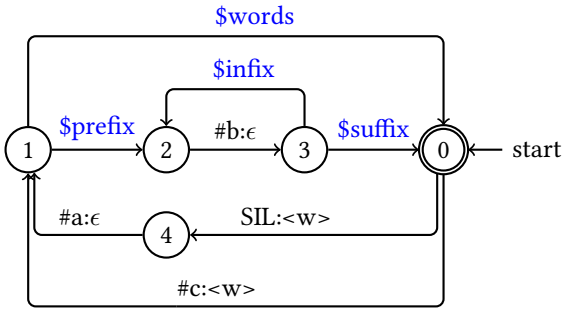

Figure 2: Prototype Lexicon FST for subword-based lexicon.

are shown. Note that the actual realization of the boundary character (here a +-sign) does not matter, but the locations of these markers do.

Style (abbreviation)	Example
Boundary Tag(<w>)	<w> dan <w>rádje riikka t <w>
left-marked (+m)	dan rádje +riikka +t
right-marked (m+)	dan rádje+ riikka+ t
left+right-marked (+m+)	dan rádje+ +riikka+ +t

Table 1: Four methods to mark the subword units in the sequence "dan rádjeriikkat"

As the n-gram language models are trained on the subword units, high-order n-grams are needed to provide a context of a reasonable length. We use the Kneser-Ney growing algorithm (Siivola et al., 2007) to train high-order Kneser-Ney smoothed varigram models.

2.3 Deep neural networks

We experiment with three different neural network architectures, all of which have demonstrated the ability to model speech well with large amounts of data.

A time delay neural network (TDNN, Peddinti et al., 2015) is a type of a feedforward network. The main benefit for speech recognition is modeling the changes in duration

and varying boundaries of phonemes in the speech signal. It is constructed by having also a time delayed copy of the signal as an input. This helps the network to disregard varying start and end points of the pattern in its classification.

TDNN models can be improved by using different training criteria that match the task of speech recognition better. Regular TDNN models are trained on a frame-based cross-entropy criterion. This means that the recognizer optimizes for the recognition of phones in each separate frame. Although this sounds ideal and works well in practice, it can be further improved upon by using a criterion that actually looks to the power to predict a sequence of phones. In Povey et al. (2016) these models are introduced and named "Lattice-free maximum mutual information" or colloquially "chain models". During the training of the network, a window of frames is not only classified, but a simple forward-backward algorithm is run to estimate the sequence that will be predicted by the real speech recognizer.

Long short-term memory (LSTM) networks are a variant of recurrent neural networks (RNN). In basic RNNs the state of the hidden layer is fed back to the next step as one of the inputs, giving the network a memory of the previous inputs. However, having many hidden layers might lead to a vanishing gradient problem, where during training the gradient "vanishes" while it propagates back in the network. To correct for this, LSTMs use a so-called memory cell, to balance which information should be carried for multiple steps in the network in "long-term memory", and when to use this information in the calculations for the current state in "short term". For a bidirectional-LSTM (BLSTM), this is happening in both directions.

3 Experiments

We start by demonstrating the improvements obtained without DNNs by Kaldi and WFST-based decoding in relation to the AaltoASR and token-passing decoding. We continue by comparing different subword boundary markings and choose the overall best for the next experiments, where we compare different types of DNN architectures for acoustic modeling. Finally, we show the effects of increasing the size of the language model training data.

3.1 Data

We use the same data sets as Smit et al. (2016) to provide a fair comparison. The data includes audio data from the UIT-SME-TTS corpus with one female and male speaker. For both speakers we train a speaker-dependent recognizer using 2.5 hours of audio. Rest of the data is divided into development and evaluation sets 3:2, roughly 1–1.5 hours total. Our initial language models are based on 10 000 randomly selected sentences from the Northern Sámi Wikipedia dump in addition to the acoustic model training sentences (TRAIN+WIKI). Further tests with a larger corpus are based on "Den samiske textbanken" (BIG).

Audio

Speaker	Gender	Title	Amount
SF1	Female	UIT-SME-TTSF	3.3 hours
SM1	Male	UIT-SME-TTSM	4.6 hours

Text

Source	# sentences	# word tokens	# word types
Sami Wikipedia	10k	88k	20k
Den samiske textbanken	990k	12M	475k

Table 2: Language and acoustic modeling data for the speech recognizer training.

3.2 Setup

We started by first building a simple monophone-based model on MFCCs extracted from the training data and used this to better align our audio data to the transcript. After this step, we trained a traditional triphone GMM-HMM model on these improved alignments.

For our TDNN we iterate the previous step by again aligning our data with the GMM-HMM model and used these alignments together with speed and volume perturbated training data for higher dimensional MFCC features. As a result, we get a five layers deep TDNN. A similar process was used to train the BLSTM and Chain model to generate networks with seven and six layers respectively.

For a word-based system, we trained a Kneser-Ney smoothed 3-gram model with the SRILM toolkit (Stolcke, 2002). For subword language modeling, we first trained a Morfessor model based on the TRAIN+WIKI corpus. We used Morfessor 2.0 implementation (Virpioja et al., 2013) with token-based training and the corpus weight parameter as 1.5. The words in the corpus were segmented to subword units with the aforementioned model using each of the different subword boundary markings. The subword n-gram models were then trained on the corpora using the VariKN toolkit (Siivola et al., 2007) with maximum n-gram length as 10.

For the BIG corpus we trained both 3-gram and 10-gram models with the same tools. The smaller model was used for first pass scoring and 10-gram model used afterward to rescore the lattices. In TRAIN+WIKI all results are with a single-pass 10-gram model. Table 3 shows the size of the different language models (LM) and lexicons. The ASR lexicon size varies due to the different subword boundary markings even if the words are segmented with the same Morfessor model.

We report for all experiments both the word error rate (WER) as well as the letter error rate (LER). The former is more common in general speech recognition research, while the latter is more common in evaluating speech recognition for agglutinative languages, where minor mistakes such as selecting a wrong inflectional suffix or splitting a compound word have very strong effects on WER.

3.3 Results

Table 4 compares the error rates of the GMM-HMM baselines from AaltoASR and Kaldi. Since the data and language models are the same the difference is due to the

Data	Units	Lexicon (#types)		LM (#n-grams)	
		SF1	SM1	SF1	SM1
TRAIN+WIKI	words	23.5k	23.1k	103.9k	102.4k
	subwords, <w>	14.3k	14.1k	751.8k	747.6k
	subwords, +m+	19.1k	18.7k	610.9k	600.0k
	subwords, +m	16.1k	15.8k	608.7k	596.9k
	subwords, m+	17.2k	17.0k	607.5k	596.4k
BIG	words	474.9k		5.9M	
	subwords, <w>	93.9k		51.6M	
	subwords, +m+	172.4k		64.6M	
	subwords, +m	122.2k		65.0M	
	subwords, m+	137.8k		64.4M	

Table 3: Lexicon and language model sizes for word models and subword models with different boundary marking styles.

	SF1		SM1	
Toolkit	WER	LER	WER	LER
AaltoASR	37.5	8.5	39.5	9.4
Kaldi	32.3	6.9	34.9	7.4

Table 4: Comparison between AaltoASR (Smit et al., 2016) and Kaldi with 10-gram LM based on TRAIN+WIKI and 2.5h of audio for both speakers.

toolkits, the decoders, and the GMM-HMMs implementations.

Table 5 continues with the Kaldi system to compare the four subword boundary markings. The differences are small given the size of the test data, but the traditional word boundary tag <w> seems to be a good choice and was used in the further experiments. It has the smallest lexicon, but because the boundary tag consumes one position in each n-gram context longer n-grams are utilized than in the other models. However, because the subword LMs are trained with the VariKN toolkit, the increase in the LM size is minimal.

	SF1		SM1	
Language Model	WER	LER	WER	LER
word 3-gram	43.9	9.2	49.7	10.4
subword 10-gram, <w>	32.3	6.9	34.9	7.4
subword 10-gram, +m+	33.8	7.1	38.1	8.2
subword 10-gram, +m	32.5	6.9	36.2	7.5
subword 10-gram, m+	36.5	7.0	38.9	7.4

Table 5: Error Rates for different subword boundary markings. All models were trained with the TRAIN+WIKI corpus and 2.5h of audio.

Table 6 presents the main result of this paper, which is the comparison of GMM-HMM to various DNN architectures when the training data resources are limited. The special advantage of DNNs is their remarkable effectiveness in modeling ”deep”

Speaker	Acoustic model		TRAIN+WIKI		BIG	
	Type	#params	WER	LER	WER	LER
SF1	AaltoASR	600k	37.5	8.5	23.7	5.5
	HMM-GMM	858k	32.3	6.9	19.9	3.8
	TDNN	6.6M	24.8	4.9	14.7	2.5
	Chain Model	5.8M	25.6	6.0	17.0	3.5
	BLSTM	10.8M	25.6	5.3	13.9	2.7
SM1	AaltoASR	600k	39.5	9.4	20.9	4.9
	HMM-GMM	858k	34.9	7.4	18.0	3.6
	TDNN	6.6M	29.2	5.7	12.5	2.1
	Chain Model	5.8M	29.8	6.0	15.2	2.8
	BLSTM	10.8M	28.5	5.8	12.8	2.4

Table 6: Error Rates between TRAIN+WIKI and the BIG language model. Same acoustic data was used in all models. AaltoASR results are from Smit et al. (2016).

structures in data that the previous frameworks could not take into account. In speech recognition, this has been taken to mean that DNNs require large amounts of training data. However, it is possible that in limited applications such as speaker-dependent systems, DNNs may be able to find useful structures even from small amounts of data. Table 6 shows clear improvements in every DNN architecture compared to the GMM-HMM method. At the point of writing, our simplest network TDNN is at least as good or better than the more complex Chain model and BLSTM, but given more time to study optimal hyperparameters for small data settings, we might be able to train models surpassing the now new baseline.

Finally, Table 7 shows that the relative differences between different subword boundary markings do not change much even when the language models are trained using the larger corpus. As in Table 5, the relative differences are small given the size of the test data, but the traditional word boundary tag <w> is still unbeaten and all subword models are better than the word-based model.

Language Model	SF1		SM1	
	WER	LER	WER	LER
word 3-gram	17.6	3.1	17.0	2.8
subword 10-gram, <w>	14.7	2.5	12.5	2.1
subword 10-gram, +m+	14.9	2.8	13.4	2.3
subword 10-gram, +m	14.6	2.7	14.6	2.4
subword 10-gram, m+	16.3	2.6	13.7	2.3

Table 7: Error Rates between different boundary marking styles using the BIG language model. TDNN was used in all recognizers.

4 Conclusions

In this paper, we applied the state-of-the-art ASR framework based on Kaldi and DNN acoustic models to get a new baseline for Northern Sámi. The results were quite im-

pressive with up to 50% relative error rate reduction. The only drawback in WFST-based speech recognition with large LMs is the size of the WFST search graph, which makes the memory consumption of the single pass decoding sometimes prohibitive. However, in most cases this can be compensated by a two-pass recognition where the second pass is used to rescore the existing search graph with the large LM. The single pass approach does also provide reasonable results already with a low order n-gram models. In addition, the modeling of position-dependent phones and other advanced acoustic modeling developments implemented in Kaldi was a clear benefit. Considering these it is recommended to apply Kaldi for the following research.

The results show clearly that at least in speaker-dependent systems, even with relatively small amounts of audio data, the DNNs were capable of finding structures in data that made them superior to the old state-of-the-art GMM-HMM models. DNNs are also very complex, and their techniques and methods are continuously advancing, so we expect to still achieve further significant improvements in near future. Also, even with the current techniques we should be able to improve the results further by more thoroughly optimizing the layer sizes and hyperparameters of the neural networks. For example, Mansikkaniemi et al. (2017) was able to improve the state-of-the art results for Finnish broadcast news results by 3% relative with such optimizations.

For the different types of subword boundary markings, our experiments resulted only small differences for Northern Sámi. Although the traditional word boundary tags gave slightly better results than the other marking styles more studies should be performed on how much the results depends on the language, data, and the length of the subword units.

The next step for improving the LMs in Northern Sámi is to apply recurrent neural networks. For RNNLMs, the whole word units have further disadvantages in morphologically rich languages, because the large vocabulary increases the dimensions of the input and output layers. For Finnish, using RNN language models with subword units has lowered the WER by 11% with a large training corpus (Smit et al., 2017a). Reducing the corpus size from 160 million tokens to 16 million tokens, which is close to our BIG data set for Northern Sámi, reduced the improvement only slightly to 9%. Smit et al. (2017a) show also promising results for Finnish and Arabic with purely character-based models.

For under-resourced languages specifically, an interesting future direction is to develop methods to better take advantage of a well-resourced related language. Even simple methods such as data pooling, acoustic model adaptation or bootstrapping with large amounts of unlabeled data have been popular. For Northern Sámi we could, for example, try to apply the data and expertise available in Finnish and Estonian. Regardless of the approach taken to improve the ASR, the system build in this paper provides a good baseline for further experiments.

Acknowledgements

We thank the University of Tromsø for the access to their Northern Sámi datasets and acknowledge the computational resources provided by the Aalto Science-IT project.

This work was financially supported by the Tekes Challenge Finland project TELLme, Academy of Finland under the grant number 251170, and Kone foundation.

References

Ebru Arisoy, Dogan Can, Siddika Parlak, Hasim Sak, and Murat Saraclar. 2009. Turkish broadcast news transcription and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing* 17(5):874–883. https://doi.org/10.1109/TASL.2008.2012313.

Ghinwa Choueiter, Daniel Povey, Stanley F. Chen, and Geoffrey Zweig. 2006. Morpheme-based language modeling for Arabic LVCSR. In *ICASSP 2006 – IEEE International Conference on Acoustics, Speech and Signal Processing*. pages 1053–1056. https://doi.org/10.1109/ICASSP.2006.1660205.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL 2002 Workshop on Morphological and Phonological Learning*. Association for Computational Linguistics, Stroudsburg, PA, USA, volume 6 of *MPL '02*, pages 21–30. https://doi.org/10.3115/1118647.1118650.

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

Geoffrey Hinton, Li Deng, Dong Yu, Goerge E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* 29(6):82–97. https://doi.org/10.1109/MSP.2012.2205597.

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

Teemu Hirsimäki, Janne Pylkkönen, and Mikko Kurimo. 2009. Importance of high-order n-gram models in morph-based speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 17(4):724–732. https://doi.org/10.1109/TASL.2008.2012323.

André Mansikkaniemi, Peter Smit, and Mikko Kurimo. 2017. Automatic construction of the Finnish Parliament Speech Corpus. In *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*. Stockholm, Sweden.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2008. Speech recognition with weighted finite-state transducers. In *Springer Handbook of Speech Processing*, Springer, pages 559–584.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*. Dresden, Germany, pages 3214–3218.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan

Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *ASRU 2011 – IEEE Workshop on Automatic Speech Recognition & Understanding*.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*. San Francisco, pages 2751–2755. https://doi.org/10.21437/Interspeech.2016-595.

Janne Pylkkönen. 2005. 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.

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

Peter Smit, Siva Reddy Gangireddy, Seppo Enarvi, Sami Virpioja, and Mikko Kurimo. 2017a. Character-based units for unlimited vocabulary continuous speech recognition. In *ASRU 2017 – IEEE Workshop on Automatic Speech Recognition & Understanding*.

Peter Smit, Juho Leinonen, Kristiina Jokinen, and Mikko Kurimo. 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*. pages 80–91.

Peter Smit, Sami Virpioja, and Mikko Kurimo. 2017b. Improved subword modeling for WFST-based speech recognition. In *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*. Stockholm, Sweden.

Andreas Stolcke. 2002. SRILM — an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-2002)*. pages 901–904.

Balázs Tarján, Tibor Fegyó, and Péter Mihajlik. 2014. A bilingual study on the prediction of morph-based improvement. In *Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014)*. pages 131–138.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Report 25/2013 in Aalto University publication series SCIENCE + TECHNOLOGY, Department of Signal Processing and Acoustics, Aalto University, Helsinki, Finland.