Computer Speech & Language 66 (2021) 101158

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Advances in subword-based HMM-DNN speech recognition across languages

Peter Smit*[a,c], Sami Virpioja[b,a,d], Mikko Kurimo[a]

[a] Department of Signal Processing and Acoustics, Aalto University, Finland
[b] Department of Digital Humanities, University of Helsinki, Finland
[c] Inscripta, Finland
[d] Utopia Analytics, Finland

ARTICLE INFO

Article History:
Received 8 January 2020
Revised 26 June 2020
Accepted 13 September 2020
Available online 28 September 2020

Keywords:
Large vocabulary speech recognition
Subword units
Character units
Recurrent neural network language models

ABSTRACT

We describe a novel way to implement subword language models in speech recognition systems based on weighted finite state transducers, hidden Markov models, and deep neural networks. The acoustic models are built on graphemes in a way that no pronunciation dictionaries are needed, and they can be used together with any type of subword language model, including character models. The advantages of short subword units are good lexical coverage, reduced data sparsity, and avoiding vocabulary mismatches in adaptation. Moreover, constructing neural network language models (NNLMs) is more practical, because the input and output layers are small. We also propose methods for combining the benefits of different types of language model units by reconstructing and combining the recognition lattices. We present an extensive evaluation of various subword units on speech datasets of four languages: Finnish, Swedish, Arabic, and English. The results show that the benefits of short subwords are even more consistent with NNLMs than with traditional n-gram language models. Combination across different acoustic models and language models with various units improve the results further. For all the four datasets we obtain the best results published so far. Our approach performs well even for English, where the phoneme-based acoustic models and word-based language models typically dominate: The phoneme-based baseline performance can be reached and improved by 4% using graphemes only when several grapheme-based models are combined. Furthermore, combining both grapheme and phoneme models yields the state-of-the-art error rate of 15.9% for the MGB 2018 dev17b test. For all four languages we also show that the language models perform reasonably well when only limited training data is available.

1. Introduction

The term large vocabulary continuous speech recognition typically refers to speech recognizers operating on tasks that have a broad domain and cover the majority of regularly used words. Implicitly, it also assumes that a speech recognition system has a vocabulary which must be large to sufficiently cover the language used. Originally, the term also implies that the vocabulary consists of words, similarly as how the vocabulary is treated in linguistics, except that often the surface forms of the words are used explicitly.

*Corresponding author.
E-mail address: peter.smit@inscripta.io (P. Smit).

https://doi.org/10.1016/j.csl.2020.101158

Unfortunately, even the largest vocabularies limit the use of words into the listed ones. However, it is not possible to create an exhaustive list of words, because all training data is also limited and languages evolve with new words appearing and old words being forgotten. In speech recognition, the coverage of the vocabulary is often measured by the out-of-vocabulary rate. It is the proportion of words that are not present in the vocabulary and thus cannot be recognized correctly.

For morphologically rich languages, such as Finnish, attempts to make a vocabulary to cover a sufficient part of the language is even more problematic. Word formation by derivation and compounding produces a massive number of lexical words, and inflectional processes further create a large number of variations requiring dozens or more surface forms to cover a single lexical word.

In the context of these problems, subword-based speech recognizers have been developed to cope with unlimited (or open) vocabulary tasks (Bisani and Ney, 2005; Hirsimäki et al., 2006). Instead of words, the vocabulary contains morphemes, syllables, or other subword units that together can be used to create an unlimited amount of word forms. If the units have been appropriately chosen, all words in the language can be generated and modeled by the system. This would include words not seen in the training data or even words that might not even have existed yet at the time the system is created. One step further is to create a vocabulary-free system based only on characters. These systems have full freedom to predict any word, as long as it can be written by the characters known to the system.

A subword or vocabulary-free system also has drawbacks. Some languages, such as English, have strong pronunciation variations in words and character sequences, which are hard to model on the subword or character level. While this could be solved by finding and introducing the relevant pronunciation variants in the pronunciation dictionary for subwords, we take here a phoneme-free approach and build the acoustic models directly for graphemes. Another drawback of subword lexicons is that while they allow producing proper words that were not seen in the training data, they can similarly generate non-words. During decoding, the possible search space is larger than for word-based recognition, and language models that can use long subword contexts are needed to guide the decoder.

Besides in Hybrid HMM-DNN based systems, the character and subword-based language models are also relevant for ‘end-to-end‘ trained models such as Graves and Jaitly (2014); Chan et al. (2016). Although the latest end-to-end systems perform very well, the Hybrid HMM-DNN systems are often still the preferred choice. This is, for example, the case when separate language models are needed, or when much more text data is available than audio data. Especially in low-resource situations this is often the case. However, the end-to-end systems do use subword and charater-based models for much of the same reasons, including the size of layers in the neural networks as well as the more effective use of training data. Like in our systems, the grapheme-based acoustic models are also relevant for end-to-end models and often are the preferred choice (Sainath et al., 2018; Rao and Sak, 2017).

In the following sections, we describe how we created effective subword and character based models in a conventional speech recognition system, where the acoustic and language models are trained separately. We introduce the concepts needed to implement subwords correctly in a weighted finite state transducer (WFST) based recognizer and the considerations needed to train both conventional *n*-gram language models as well as modern recurrent neural network language models. Besides merely testing systems with different units, we also propose, implement, and evaluate system combinations that combine models composed of different units.

The experiments are repeated on benchmarks covering four different languages from three different language families with different vocabulary sizes and morphological complexities. For all of these four speech datasets we obtain, to the best of our knowledge, the best results published so far. The motivation for this work came from the improvements we obtained by developing the new subword WFSTs (Smit et al., 2017c), character-based language models (Smit et al., 2017b) and then winning the 2017 multi-genre broadcast speech recognition challenge by combining systems operating on different subwords (Smit et al., 2017a). While some of the ideas were initially presented in these conference papers, they have now been extended and analyzed in detail here. All the results are entirely new and the tasks include also Swedish and English, for which such methods have not been proposed before.

In summary, the main novelty in this work is the set of tools and techniques for successful subword modeling for WFST-based hybrid DNN-HMM speech recognition using graphemes instead of phonemes. In addition, it includes a thorough evaluation across four diverse languages as well as an evaluation of these techniques in an under-resourced scenario. Moreover, it explores the usage and considerations in using subword and character based neural-network language models for hybrid DNN-HMM systems. Lastly, it introduces and evaluates the tools for doing lattice combination across different language modeling units, allowing, for example, Minimum Bayes Risk decoding with a larger variety of models.

2. Subword modeling for ASR

Subword modeling has been used for almost twenty years in speech recognition. For some languages, such as Arabic, it has been popular to use linguistic units such as morphemes as the basic language modeling unit (Choueiter et al., 2006; Kirchhoff et al., 2006; Mousa et al., 2013). For other languages, such as Finnish, the subword segments are often created with data-driven methods (Hirsimäki et al., 2006; Creutz et al., 2007). Multiple data-driven methods for subword segmentation have been used in speech recognition. In Smit et al. (2017c) we have tested multiple methods such as byte-pair encoding (Gage, 1994), Morfessor (Creutz and Lagus, 2002) and Greedy Unigram (Varjokallio et al., 2013). The results show that if the parameters and context length for the language model are optimized to a comparable level, the actual segmentation method has only a minor effect on the speech recognition performance.

Not only do these systems need a subword segmentation, but also a method to reconstruct words back from the subword units. This can be done by e.g. adding a dummy unit to mark a word boundary, or by creating different subword variants based on the location in a word. The first subword systems often used a separate word boundary marker (Hirsimäki et al., 2009) or a continuation marker attached on the left side of the subword when there is no word boundary (Arisoy et al., 2009; Tarján et al., 2014). In Smit et al. (2017c) we showed that the selected marking style can actually have a profound effect on the speech recognition result and that the optimal marking style often depends on the data set. Table 1 shows a list of the common marking styles, as well as the corresponding abbreviations used in this paper.

After segmenting the training texts into subwords, conventional tools can be used for training the language models by treating the marked subwords as independent words. Only when the word-based perplexity is calculated, the actual reconstructed words need to be accounted for. If the word boundary tags (<w>) are used for that, these tags will be treated as normal tokens which have their own probability. This is necessary to predict the word boundaries correctly, otherwise any location of word boundary tags would be as likely.

In this work we design our subword modeling in such way that it is independent from the acoustic model. The acoustic model can be trained on sequences of phones (or graphemes), and any type of language model can be used with it, whether it is using characters, other subwords, or words.

2.1. Data-driven subword segmentation

Multicharacter subword units are typically derived in an unsupervised manner from the training data. Smit et al. (2017c) compared three data-driven segmentation methods for Finnish ASR: Morfessor Baseline (Creutz and Lagus, 2002; 2007; Virpioja et al., 2013), Greedy Unigram (Varjokallio et al., 2013), and the byte-pair ecoding (BPE) algorithm popular in neural machine translation (Gage, 1994; Sennrich et al., 2015). The Morfessor segmentation was found to be slightly but consistently better on three different lexicon sizes, so we use it in this work.

Morfessor Baseline defines a unigram model over the subword segments. The parameters (i.e. the segments) are optimized for maximum a posteriori criterion, with a prior inspired by the Minimum Description Length principle (Rissanen, 1978). A greedy local search algorithm is used for minimizing the cost function (see, e.g., Virpioja et al., 2013). The granularity of segmentation can be controlled by adding a hyper-parameter α between the cost of encoding the lexicon and the cost of encoding the corpus part in the cost function:

$$L(\theta, D_W) = -\log P(\theta) - \alpha \log P(D_W|\theta), \tag{1}$$

where θ includes the model parameters and D_W are the words in the training data. A small value of α emphasizes the prior, resulting in shorter subwords and smaller lexicon than with a large value. The training can be based on word tokens, word types, or using the word frequencies with logarithmic dampening (Virpioja et al., 2011).

2.2. Pronunciation lexicon

In a traditional speech recognition system, the vocabulary is directly linked with the pronunciation lexicon. For languages such as English, where the pronunciation is not based on rules, often hand-crafted lexicons that map each word to the correct phoneme sequence are used. Typically, most words include only a single or only a few pronunciations.

When words are split into subwords, it is not always clear what the pronunciation of each subword is. A word can have more or fewer phonemes than graphemes, and the relation that might be obvious to humans might be hard to determine algorithmically. Many subwords will also have multiple pronunciation variants. Moreover, many subwords will share the same pronunciation sequence, which increases the complexity of the recognition process.

One solution to this problem is to use grapheme-based units instead of phonemes, and let the acoustic model learn the pronunciation patterns from the training data. In languages that are highly phonemic, e.g. Finnish, this is a natural thing to do. For languages that have a more irregular pronunciation pattern, e.g. English, this is more complicated, because the acoustic model must learn the pronunciation of characters based on their context. In Gaussian mixture model (GMM) based acoustic models, the phoneme-based recognizers outperform the grapheme-based ones for English by margins as big as a 50% relative increase in word error rate (WER) (Mirjam Killer, 2003). However, a recent work (Wang et al., 2018) shows that modern deep neural

Table 1
Four methods of marking subword units so that the original word sequence ‘two slippers‘ can be reconstructed.

Style (abbreviation)	Example
boundary tag (<w>)	<w> two <w> slipp er s <w>
left-marked (+m)	two slipp +er +s
right-marked (m+)	two slipp+ er+ s
left+right-marked (+m+)	two slipp+ +er+ +s

network (DNN) acoustic models can learn the English pronunciations better, and the difference between phoneme and grapheme-based recognizers can be as low as 5% relative WER. Also, the modern sequence-to-sequence trained recognizers often predict graphemes directly without using a separate pronunciation lexicon (Prabhavalkar et al., 2017).

In this work, we used grapheme-based acoustic models for all languages, which solves the problem of generating pronunciations for all subwords.

2.3. Subword modeling in WFST based speech recognition

In the weighted finite state transducer (WFST) framework (Mohri et al., 2008) a decoding graph is created by the composition of four different FSTs, abbreviated with the term HCLG. The H-FST maps HMM states to context-dependent phones, the C-FST maps these to context independent phones. The L-FST (lexicon) maps the phone sequences to words and the G-FST scores the sequences of words with a language model.

In a basic implementation of the decoding graph, it would be possible to use subwords by simply using those instead of words when creating the separate FSTs. This would naturally still need a method for reconstructing words, but this could be handled by using markers on the subwords, similar as has been done in this paper. However, there are two common extensions made to the lexicon FST, which do have an impact on subword models. First, the lexicon FST often allows optional silence phones to be inserted between words and on the beginning and end of the utterance. However, in subword FSTs, there cannot be any silence phones between subwords that belong to the same word. Second, in words, different phoneme-variations are used depending on the location, e.g. whether the phoneme is the first or last one in a word. To take this into account, also the subword FSTs must be aware of their position in the current word. For example, when using the +m+ marking style with the subwords 'hel+' and '+lo' we want 'hel+' to map to the phones 'h_B e_I l_I' where '_B' indicates a phone on the beginning of a word and '_I' a word-internal phone. Similarly, for '+lo' we want the phones 'l_I o_E' where '_E' indicates a phone on the end of a word.

Fig. 1 shows how the lexicon FST can be created in such a way that these properties still hold for subword-based systems. The subword lexicon is split into four categories (prefix, infix, suffix, and complete words) in which the position-dependent phones are marked appropriately. Note that depending on the marking style a subword might appear in multiple categories. E.g., with the <w> -style, all subwords belong to all four categories. More details and evaluations can be found in Smit et al. (2017c).

2.4. Language modeling with subword units

The choice of language modeling unit has significant implications for a speech recognizer. Although the tools and techniques do not necessarily change—simple *n*-gram models will work to a certain degree—the optimization of these tools and techniques is very dependent on the chosen unit. The choice of units changes the characteristics of the tokenized training text, changing both the number of different tokens, the number of tokens in the text, and the number of times a token is used. When words are split more, the number of tokens increases and number of types decreases. The extreme case is to split into characters, where the lexicon would constitute of only the character set.

When the number of tokens in a sentence increases, it increases the number of units needed to represent the required context information. In *n*-gram models, the only way to capture the more context information with shorter units is to increase the *n*, the length of the preceding context. Without any pruning techniques, this would also increase the size of the model exponentially. For this reason, we use the VariKN toolkit (Siivola et al., 2007), which uses dynamic modified Kneser-Ney growing and pruning of *n*-gram models. The resulting models, so-called varigrams, do not fix the length of the preceding contexts but let it depend on the improvement it gives to the model. In practice, a model with characters as units might have the majority of *n*-grams with a context between 8–14 preceding tokens. The algorithms implemented in VariKN have been compared to traditional fixed-length *n*-gram models for Finnish and Estonian (Hirsimäki et al., 2009) and Northern Sámi (Smit et al., 2016) speech recognition, showing clear improvements. Fig. 3 in Section 3.1 illustrates the distribution of *n*-grams used by some of the varigram models trained in our experiments.

For recurrent neural network (RNN) based language models, it is of less importance how many tokens there are in a sentence, as the models can learn dependencies that occur between tokens further apart in history (Mikolov et al., 2010). Besides that, the subword models have an advantage over word models because of their smaller vocabulary size. If the vocabulary of a system is too large, a lot of computing and training time is used for learning the parameters of the input and output layer which contain the same number of units as the vocabulary size. Multiple methods to combat this, such as class-based training (Botros et al., 2015) or hierarchical softmax (Morin and Bengio, 2005; Kuo et al., 2012) have been proposed, but for the subword models, these methods are typically not needed as the size of the input and output layer is reasonable by default.

For both *n*-gram and RNN language models, the most prominent advantage can be found in the increased vocabulary coverage and the reduced sparsity of the training data. The data-driven subword models mentioned in Section 2.1 have the ability to predict almost any word in the language, even words that were not seen in the training data and those that only came into existence after the models were trained. Also, the units of subword models occur more frequently and in more different contexts, giving the models more examples to train from. In a word-based system, more than half of the words might only appear once or twice, which is not enough to build a robust estimate of their occurrence in future texts.

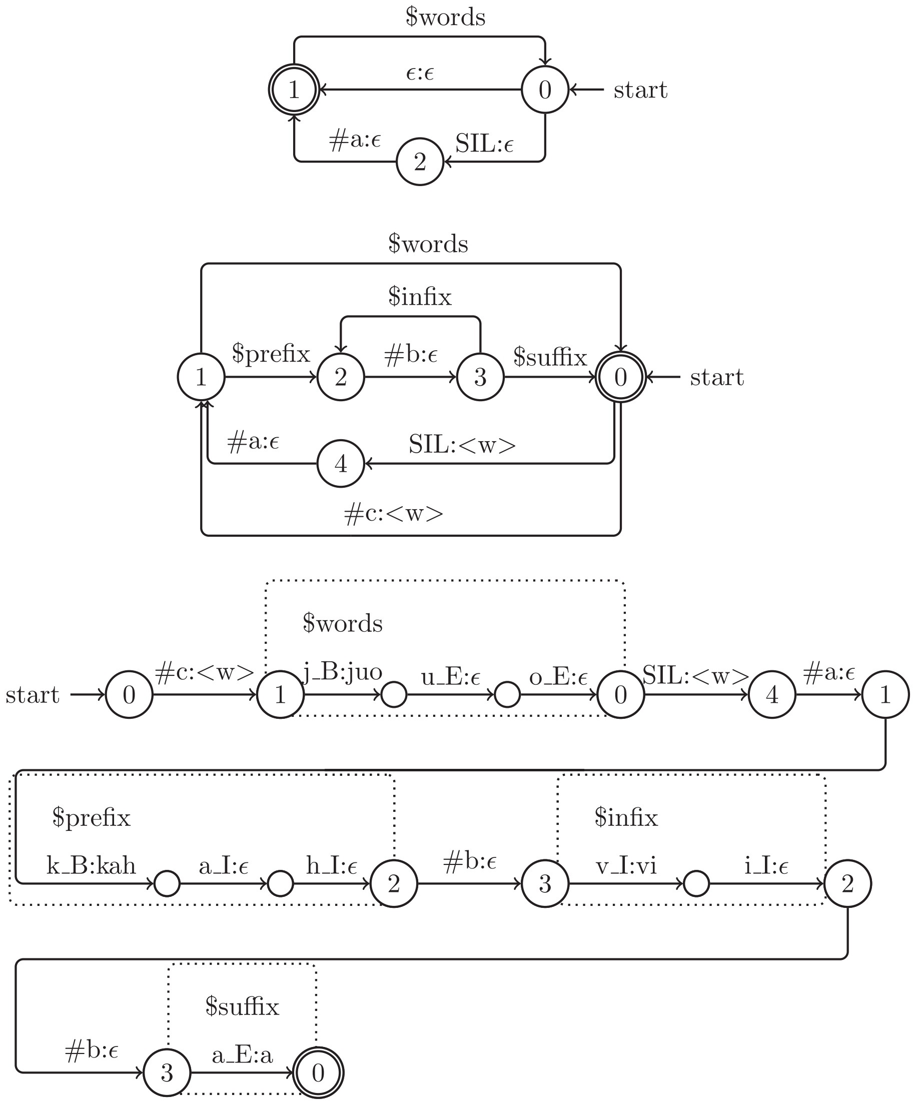

Fig. 1. The original lexicon FST (top) and the subword-modified lexicon FST (middle). The labels on the arcs show the input:output label for each transition. ϵ represents the empty symbol. Labels starting with '#' are dummy labels that are used for keeping the FSTs deterministic (a requirement in the Kaldi toolkit). Labels that start with $ are replaced with linear FSTs containing the actual pronunciations of the applicable words. The bottom figure is an expanded example FST for the sequence ' <w> juo <w> kah vi a' (translation: 'drink coffee')

3. System evaluations

To evaluate the properties of subword and character-based models in a modern neural network-based speech recognition system, we have built systems for benchmarking tasks in four different languages from three different language families. First, we evaluate for each language the general performance of a word-based model and then the performance on different subword marking styles for *n*-gram-based language models. After that, we select the best marking and evaluate the different levels of segmentation by changing the corresponding parameter in Morfessor. We evaluate these models both on *n*-gram language models and two different recurrent neural network architectures; one "shallow" and one "deep" neural network. The details of these systems are described in Section 3.1.

3.1. Setup

We have used the same set of tools and recipes to create the acoustic and language models for all four languages.

3.1.1. *Acoustic model training*

The acoustic models are trained with the Kaldi toolkit (Povey et al., 2011). First GMM-based models are trained, which are used to automatically clean and segment the data with the standard facilities present in Kaldi. The resulting cleaned data set is used to train a neural network based acoustic model. For all languages, we trained three different models. The most basic one is a time-delay neural network (TDNN) (Peddinti et al., 2015), which is a non-recurrent neural network that takes a fixed time window as input. Furthermore, the parameters of lower layers are shared between smaller time windows in a similar fashion to convolutional neural networks. The second NN-based model was a mixture of TDNN layers and long short-term memory layers (LSTM), which is a recurrent architecture that retains information from previous samples to improve its modeling power. The last model is a bi-directional TDNN-LSTM model, which also unrolls a network forward in time (Cheng et al., 2017). In our experience, the recurrent models require more data in order to be trained adequately. Therefore, they are not expected to outperform the regular TDNN model for smaller data sets.

All neural network acoustic models are trained with lattice-free MMI (Povey et al., 2016), which does not primarily optimize frame-based phone prediction accuracy, but instead decodes part of the utterance on the fly and calculates the error with the Maximum Mutual Information as a sequence-based criterion. However, frame-based cross-entropy is still used for regularization.

For all languages, we first used a simple word *n*-gram model to determine the best of the three optional acoustic models and then used that model through all experiments.

3.1.2. *Subword segmentations*

The different subword segmentations, besides the character one, are trained with the Morfessor 2.0 (Virpioja et al., 2013), using the basic unsupervised Morfessor Baseline algorithm (Creutz and Lagus, 2002). The training is based on word tokens. In order to obtain models that have different vocabulary sizes, multiple values for the corpus weight parameter α are used in Morfessor (Section 2.1).

The vocabulary sizes for the different languages and segmentations are shown in Table 2. In addition, Fig. 2 shows histograms for the number of units per word after segmentation. While the lexicon sizes of the largest Morfessor lexicons (with $\alpha = 0.1$) are 10−20 times higher than the smallest ($\alpha = 0.001$), they still produce a high degree of segmentation for all languages. For example, there are always more words that have been split into three units than words that are not split at all.

After segmenting the language model training text into subword tokens, four different unit boundary marker variants are created as explained in Section 2.3. For the language modeling toolkits, all variants can be trained using the same procedure; the only difference is that the marked subwords are given as input instead of words.

3.1.3. *Language models*

The *n*-gram language models are trained with the VariKN toolkit (Siivola et al., 2007). This toolkit uses modified Kneser-Ney smoothing and grows and prunes *n*-gram models dynamically, possibly including very long contexts if that results in a better model. In practice, the models are trained without limiting the order *n*. Instead, the total model size is controlled by the growing and pruning parameters. For the first recognition pass, we train a model with appr. 4−10 million *n*-gram contexts. For the rescoring pass, we train larger models, with no limits for growing or pruning. This means that all contexts that add anything to the prediction power will be added to the model. For word models, there remain typically between 50−80 million *n*-gram contexts, for character models typically up to 400 million *n*-gram contexts. The need for long contexts, especially in the case of subword models, is illustrated in Fig. 3. It shows the proportion of *n*-gram hits per order for the larger (rescoring) models on the development sets as reported by the VariKN toolkit's perplexity command. Taking Arabic as an example, in more than half of the cases the word model is able to use only 2-grams, whereas for the character model the most common hit, over 15% of all hits, is for 9-grams.

We train the RNN language models with the TheanoLM toolkit (Enarvi and Kurimo, 2016). Both a shallow and a deep version of these networks is used; their basic architecture is shown in Fig. 4. We train the models at most 15 epochs with AdaGrad, and stop early if perplexity on the development set does not decrease. For models with a very large vocabulary, we use a class-based output using classes trained with the exchange algorithm (Botros et al., 2015; Kneser and Ney, 1993). The RNN-based models are then used to rescore the *n*-gram-based lattices. Standard pruning algorithms such as limiting the amount of active search paths

Table 2
Vocabulary sizes for the different lexicons (char = characters, morf = Morfessor segmentation with the corresponding corpus weight parameter).

Segmentation	Finnish	Arabic	Swedish	English
char	29	40	32	29
morf 0.001	9 502	5 406	10 996	9 712
morf 0.01	53 183	28 802	50 537	35 441
morf 0.1	248 257	111 713	197 022	100 983
word	4 308 628	1 303 163	3 543 864	757 627

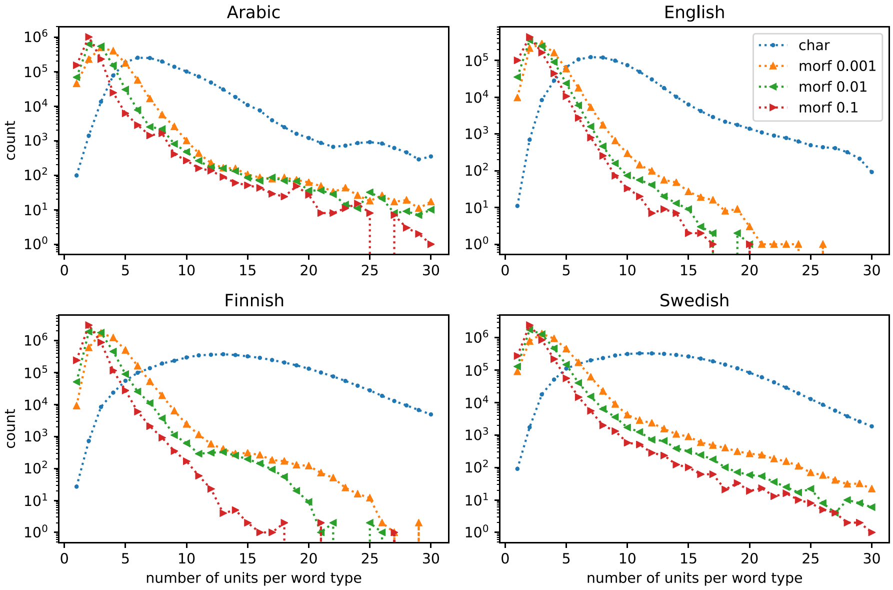

Fig. 2. Distribution of number of subwords per word for character segmentation (char) and Morfessor segmentation (morf) with different values of the corpus weight parameter.

and restricting the maximum history (from 25 tokens for word models to 100 tokens for character models) are applied in a similar manner as in Enarvi et al. (2017). The "shallow" and "deep" architectures were only mildly optimized. The "shallow" network was chosen as it corresponds to our previous work (Smit et al., 2017b) and the "deep" architecture was the same as in Enarvi et al. (2017), inspired by Srivastava et al. (2015). We further optimized the number of parameters in the deep model.

3.1.4. Evaluation scores

The language model perplexities reported in the experiments are normalized on the number of words, so that the values are comparable across different language modeling units. For models using a separate word boundary marker, the marker is considered to be part of the preceding word. Words that cannot be predicted by the model are considered out-of-vocabulary (OOV) and excluded from the perplexity values. Note that due to different out-of-vocabulary rates, the perplexities are comparable only approximatively: For a high OOV rate, many rare words are excluded, lowering the average perplexity values.

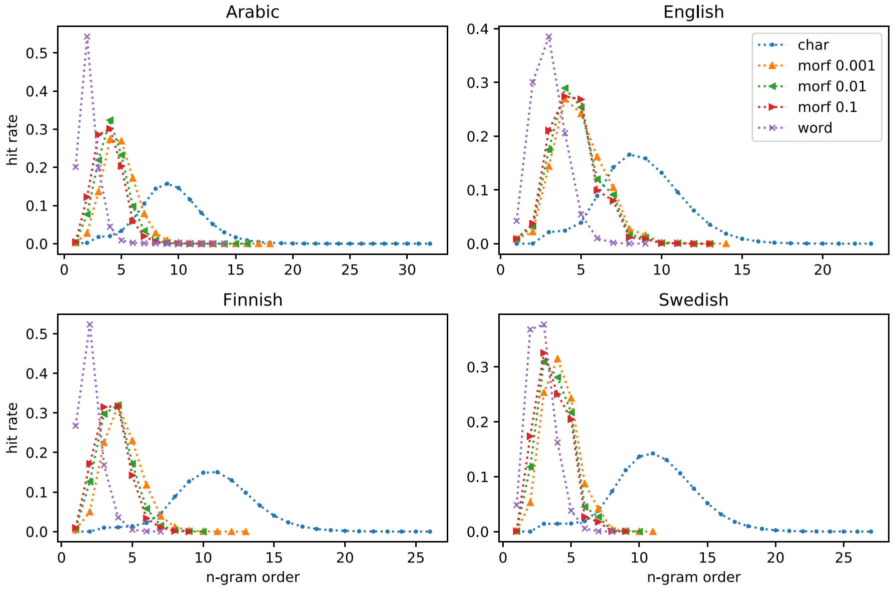

Fig. 3. N-gram hit rates for development sets using the varigram models trained for rescoring. Subword models use the <w> -style word boundary marking.

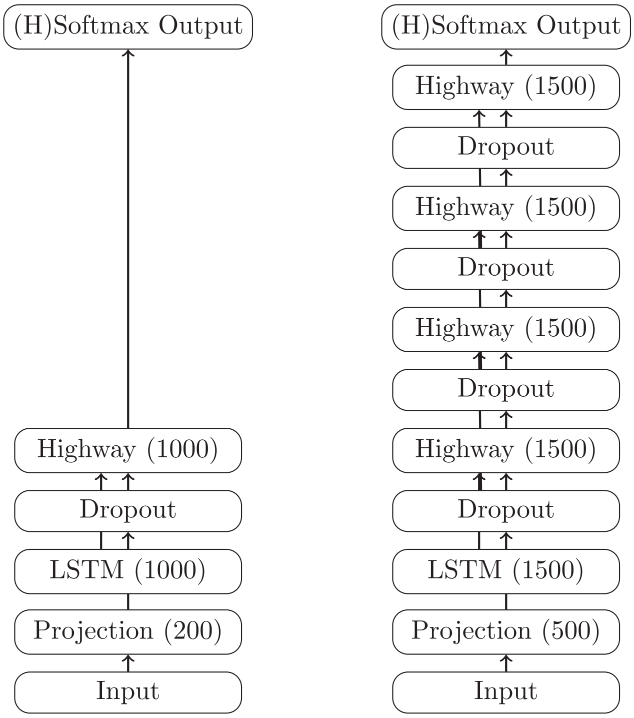

Fig. 4. The two used RNNLM architectures. Both networks start with a projection and LSTM layer (Hochreiter and Schmidhuber, 1997). On the left the ''shallow'' architecture that only contains a single highway and dropout layer. On the right the 'deep' architecture that has four pairs of dropout (Srivastava et al., 2014) and highway (Srivastava et al., 2015) layers. The number between parenthesis is the number of units in the layer.

For all speech recognition experiments, we report the word error rate (WER), which uses the counts of substitutions, deletions, and insertions to the reference text to calculate the error rate.

3.1.5. Hyper-parameter optimization

By nature speech recognition systems have many different hyper-parameters that can be tuned and chosen. Throughout this experiments we try to highlight the most important hyper-parameters in subword based systems and choose the hyper-parameters according to the results shown. Not all experiments show significant differences between multiple parameters, but we follow the local optimal result nonetheless. Naturally, from any single system it is difficult to make any conclusions that will hold across other languages or datasets. Therefore, we try to make only strong conclusions if we see a pattern that occurs across all the four languages.

3.2. Finnish

Finnish is an agglutinative language from the Uralic language family with a very large word vocabulary, which makes it especially suited to subword-based recognition. The written language is highly phonemic; a phoneme-based lexicon would be almost equivalent to a grapheme-based lexicon.

We use 1500 hours of acoustic modeling data from three different data sets. The Speecon corpus (Iskra et al., 2002) contains read speech in multiple different conditions. The Speechdat database (Rosti et al., 1998) also contains read speech from a high number of speakers over telephone lines. Lastly, the parliament corpus (Mansikkaniemi et al., 2017) is used which has speech from the Finnish parliament. For evaluation, we use a broadcast news set, obtained from the Finnish national broadcaster YLE. The test conditions are the same as in Smit et al. (2017b).

For language modeling we used data from the Finnish Text Collection (CSC - IT Center for Science, 1998) which contains mainly newspapers and books from around the turn of the century. Our selection contained 12M sentences with in total 143M tokens. The number of unique words was 4.2M.

Based on the initial experiments, where we trained three different acoustic model architectures and a small word-based *n*-gram language model, we chose the TDNN-BLSTM acoustic model for the Finnish task (see Table 3a). Note that later experiments show that the optimal acoustic model depends on the type of language model used and that in some cases a TDNN model could perform better (see Table 10a).

For Finnish, we trained two *n*-gram models for all segmentations and all different marking styles. The first-pass *n*-gram models were tuned to have appr. 4M *n*-gram-contexts. For the rescored models the amount of *n*-gram contexts depends on the level

Table 3
Evaluations for the Finnish YLE broadcast news data set.

TDNN	TDNN-LSTM	TDNN-BLSTM
19.4	18.5	**18.4**

(a) Comparison of different acoustic models using a word *n*-gram language model.

Segmentation	<w>	+m+	m+	+m
char	17.3	16.9	17.1	17.0
morf 0.001	15.9	15.7	16.2	16.4
morf 0.01	16.0	15.8	16.4	16.4
morf 0.1	16.1	15.7	16.4	16.4
word	16.7			

(b) Rescored *n*-gram results for different subword markers and segmentations using a TDNN-BLSTM acoustic model.

Segmentation	<w>	+m+	m+	+m
char	4589 (13)	4090 (35)	4158 (13)	4270 (13)
morf 0.001	3499 (13)	3507 (35)	3486 (13)	3540 (13)
morf 0.01	3489 (21)	3522 (46)	3364 (21)	3395 (23)
morf 0.1	3607 (37)	3494 (90)	3347 (48)	3390 (49)
word	2328 (810)			

(c) *n*-gram perplexity results for different subword markers and segmentations on the ASR development set. The number of OOV words are presented between parenthesis.

		RNNLM	
Segmentation	*n*-gram	shallow	deep
char	16.9	14.4	13.8
morf 0.001	15.7	13.5	13.1
morf 0.01	15.8	14.7	14.2
morf 0.1	15.7	14.2	14.0
word	16.7	15.0	14.6

(d) Comparison of different language models using a TDNN-BLSTM acoustic model and the +m+-subword marking style.

of segmentation, ranging from 50M contexts for the word model to 200M contexts for the character models. Table 3b shows that the +m+-style marker is giving the best performance, which is in line with (Smit et al., 2017c; 2017b). The Morfessor-based subword models are giving the best performance, with the best segmentation having a 5% relative improvement over a word-based model. The character model is performing similarly to the word model, having only a 1% relative degradation.

When we compare the speech recognition results with the perplexity values on the same development set (Table 3c), we see that they do not follow the same pattern. We calculated the perplexity values also for the other languages and observed similarly that they do not indicate the best subword segmentation and marking style for speech recognition. For example, the +m+-style tends to perform better in speech recognition than in language modeling, as it can help to disambiguate pronunciation variants for the same grapheme sequences.

When we rescore and interpolate the results with RNN-based language models we see a remarkable improvement for all language models. As expected, the character-based model benefits most from RNNLMs, probably because the RNN is more effective than *n*-gram in capturing long contexts. When the smaller “shallow” network and the larger “deep” network are compared, the improvement is largest for the character-based model. Compared to the previous best result in Smit et al. (2017b), 14.0%, we outperform it by 6.5% relative.

3.3. Arabic

The Arabic, as well as the Finnish language, has a structure that makes it naturally suitable for subword-based speech recognition. In previous work, linguistic units have been used frequently, but also data-driven units have been applied successfully (Choueiter et al., 2006; Creutz et al., 2007; Mousa et al., 2013).

Both the acoustic modeling and language modeling data used in this work come from the 2016 MGB-challenge. The audio is multi-genre broadcast data from Al-Jazeera and the language modeling text has been sourced from transcripts and the Al-Jazeera website. In total 1020 hours of data is used for acoustic model training and 121M tokens of text for language model training. The evaluation set used is the development set provided for the MGB2 challenge (Ali et al., 2016).

Although previously phoneme-based lexicons have shown better performance than grapheme-based lexicons (Ali et al., 2014), we have opted to use only grapheme-based lexicons to be able to run all subword systems with the same acoustic model and without preparing pronunciation dictionaries for the subwords. Note that our grapheme-based system for the MGB-3 challenge outperformed all competitors, even when only word units were used (Ali et al., 2017; Smit et al., 2017).

As for Finnish, we tested again all TDNN, TDNN-LSTM, and TDNN-BLSTM acoustic models. Table 4(a) shows that for Arabic the TDNN-BLSTM outperforms the other models by a considerable margin, hence TDNN-BLSTM models are used for further experiments.

Table 4
Evaluations for the Arabic MGB-2 broadcast data set.

TDNN	TDNN-LSTM	TDNN-BLSTM
20.8	19.4	**18.2**

(a) Comparison of different acoustic models using a word n-gram language model.

Segmentation	<w>	+m+	m+	+m
char	17.8	18.1	18.3	18.3
morf 0.001	17.3	17.2	17.2	17.4
morf 0.01	17.4	17.2	17.1	17.4
morf 0.1	17.5	17.3	17.3	17.5
word	17.7			

(b) Rescored n-gram results for different subword markers and segmentations using a TDNN-BLSTM acoustic model and the m+-subword marking style.

		RNNLM	
Segmentation	n-gram	shallow	deep
char	18.3	16.8	16.5
morf 0.001	17.2	16.0	15.7
morf 0.01	17.1	15.8	15.6
morf 0.1	17.3	16.4	16.2
word	17.7	16.7	16.5

(c) Comparison of different language models using a TDNN-BLSTM acoustic model and the m+-subword marking style.

Table 4 (b) shows the results for different segmentations and marking styles. Unlike for Finnish, none of the markings outperform the others clearly. As the best result (17.1%) was obtained with the m+-marking, we use this in further experiments. The different segmentations perform in a similar way as the Finnish ones, with the Morfessor-based models outperforming both word and character-based models.

After the RNN-based rescoring, the Morfessor segmentations are still performing the best, with the *morph 0.01* segmentation achieving a result of 15.6% word error rate. Again, the character-based model has the highest gain from increasing the depth and complexity of the RNN model. The previous best, single-system result, was 15.9% (Smit et al., 2017a).

3.4. Swedish

Swedish is a North Germanic language in the Indo-European language family. As these languages do not typically have very phonemic orthography, subword-based ASR models are not common. However, Swedish has many compound words, which suggests that subword units could work well for the lexicon.

We train our models with the data provided by the Språkbanken corpus, a public domain corpus hosted by the National Library of Norway. We used 354 hours of acoustic data from the training section of the corpus for training and 9 hours of acoustic data for development and evaluation, which is roughly 50% of the provided evaluation data.

For language modeling purposes the Språkbanken corpus contains n-gram-counts up to the 6th order, calculated from their language modeling texts. Unfortunately, the source texts, required for doing subword n-gram or RNN-based modeling, are not available. To overcome this, we reconstructed an approximation of the original language modeling corpus from the n-gram-counts that is 6-count-consistent. The procedure for this was simple, by starting with a 6-gram context that begins with a sentence marker and then obtaining the next word by finding a new 6-gram that is consistent with the last 5-gram of the current sentence. This repeats until a sentence-end marker is found and a new sentence will be started. All counts for 6-grams are decreases whenever the 6-gram is used, resulting in all 6-grams being used exactly the same amount of times as in the original language modeling corpus. Our reconstructed language modeling corpus has 398M tokens.

Although Swedish is semi-phonemic, there are enough deviations that normally a phoneme-based lexicon would be useful. Thus, the choice of implementing grapheme-based models might affect the performance. Although we have not found a comparison between phonemic and grapheme-based lexicons for Swedish, we assume based on the already excellent results showed later in this section (Table 5(c)) that a phoneme-based lexicon would not improve this system much further.

Table 5 (a) shows that the baseline for this data set already achieves a very low word error rate. In contrast to the Finnish and Arabic systems, the TDNN-LSTM system outperforms the TDNN-BLSTM system. Altough we have not investigated this in detail, we expect that the reason is the smaller amount of acoustic data in the corpus. The optimal TDNN-LSTM model had 26M parameters, while the optimal TDNN-BLSTM had already 45M parameters.

Like in the Finnish task, Table 5(b) shows that the +m+-marking style performs best for Swedish, with the <w> -style being far inferior over other styles. In n-gram-based modeling the word model slightly outperforms the Morfessor and character-based models.

Even though the word error rates for n-gram-based models are already very low, Table 5(c) shows that RNN-based language models still improve those results by a significant margin. Even more surprising is that character-based models can match the performance of word-based models. To test the diversity between these two models, we calculate the cross word error rate

Table 5
Evaluations for the Swedish Språkbanken read speech data set.

TDNN	TDNN-LSTM	TDNN-BLSTM
6.3	**6.1**	6.9

(a) Comparison of different acoustic models using a word n-gram language model.

Segmentation	<w>	+m+	m+	+m
char	3.9	3.3	3.3	3.7
morf 0.001	4.1	3.1	3.3	3.3
morf 0.01	4.3	3.1	3.2	3.2
morf 0.1	4.5	3.0	3.2	3.2
word	3.0			

(b) Rescored n-gram results for different subword markers and segmentations using a TDNN-LSTM acoustic model.

		RNNLM	
Segmentation	n-gram	shallow	deep
char	3.3	2.8	2.5
morf 0.001	3.1	2.4	2.3
morf 0.01	3.1	2.7	2.5
morf 0.1	3.0	2.6	2.4
word	3.0	2.7	2.5

(c) Comparison of different language models using a TDNN-LSTM acoustic model and the +m+-subword marking style.

(cWER) (Wong and Gales, 2017). Between the 'deep' RNN word and character model the cWER is 1.63%, showing that the majority of the errors in the models are not shared. This suggests an excellent opportunity for using these to models in an ensemble or system combination.

3.5. English

Lastly, we run the same set of experiments on an English data set. Unlike the previous languages, English has only a very limited number of inflections and surface variations for each word form, and there is a weak relation between the surface form and the pronunciation of a word. Traditionally, phoneme-based systems have always been far superior to grapheme-based system. However, it has been shown that for the 2018 MGB English data set that we are using, there is only an appr. 5% relative improvement of phoneme-based systems over grapheme-based systems (Wang et al., 2018). Therefore, even though we do not expect any necessary improvement of subword-based models over word-based models, we decided to create a grapheme-based system and evaluate the performance of word, subword and character units also in this task.

In the conditions of the official 2018 MGB challenge, the segmentation of the hour-long broadcasts needs to be done automatically. Unfortunately, we did not have a segmenter available and instead, we use the utterance level segment timings provided in the reference file. We do expect a small degradation when preparing the official results using a segmentation algorithm trained on the challenge data. In total, we used 283 hours of acoustic training data and a text corpus that consists of 646M tokens. For development and evaluation, we use the *dev17b* development set.

To validate the findings of Wang et al. (2018), we train both a grapheme and phoneme-based TDNN acoustic model for English, and compare three different word language models. Table 6 shows that in our setup the results are similar. For a simple n-gram language model, we have only a 4.3% relative WER degradation for a grapheme-based model and for the most advanced model a 6.7% relative reduction. The reason why the difference between these models is so small is not entirely evident, but we suspect that the modeling power of the acoustic models in combination with the lattice-free MMI criterion is able to compensate for the loose grapheme-phoneme relationship. However, this would require further research to be validated. As the results are very close, we are strengthened in the belief that it is possible to make competitive models for English without the use of a hand-crafted phoneme lexicon.

Table 7 (a) shows that here the TDNN (with 18M parameters) outperforms all other architectures. Given that we had the least amount of training data for this language this is not surprising and we expect that with more data to train the recurrent architectures properly they would be able to cover the wider pronunciation variation even better.

Table 7 (b) shows that for n-gram modeling the Morfessor-based models slightly outperform the word model. This is a surprising result, as previous attempts to make subword models for English have not had great success. Like the most other

Table 6
Grapheme vs. Phonemes for English using word-based language models.

	Grapheme	Phoneme
n-gram small	21.4	20.5
n-gram rescore	18.9	17.9
RNN 'deep'	17.7	16.6

Table 7
Evaluations for the English MGB dev17b broadcast data set.

TDNN	TDNN-LSTM	TDNN-BLSTM
21.4	21.6	24.2

(a) Comparison of different acoustic models using a word *n*-gram language model.

Segmentation	<w>	+m+	m+	+m
char	20.6	20.0	20.0	20.4
morf 0.001	19.4	18.7	18.8	18.9
morf 0.01	19.6	18.8	18.8	18.8
morf 0.1	19.7	18.8	18.9	18.8
word	18.9			

(b) Rescored *n*-gram results for different subword markers and segmentations using a TDNN acoustic model.

		RNNLM	
Segmentation	*n*-gram	shallow	deep
char	20.0	18.6	18.1
morf 0.001	18.7	17.9	17.8
morf 0.01	18.8	18.1	17.5
morf 0.1	18.8	17.4	17.3
word	18.9	18.3	17.7

(c) Comparison of different language models using a TDNN acoustic model and the +m+-subword marking style.

languages, the +m+-style marker performs best and is chosen for further experiments. After rescoring with an RNN-based LM we see that the subword models still outperform the word models. Table 7(c) also shows that whereas for *n*-gram models there is still a 1.1% absolute gap between word and character-based models, the gap is only 0.4% when using the deep RNNLMs.

To analyze the consistency of the English results, we have split out the test-set into the genres given by the creators of the MGB-challenge. Table 8 shows that for all categories, except the ‘advice’ genre, the Morfessor subword models outperform the word-based models. This division shows that subword models are not only applicable to very specialistic datasets, but that accross domains subwords are having benefits (or are as good as) word models.

3.6. Discussion

Looking for the common patterns in the results between languages, we first notice that the Swedish results are much better than any other tasks. A natural explanation is that it consists of read speech in a controlled environment, which is in contrast with all other data sets that have been taken from TV-broadcasts with both spontaneous speech and background noises.

Among different type of acoustic models, only those languages that had lots of data available (> 1000 hours) seem to be able to take advantage of TDNN-BLSTM models. If there are less data, the TDNN-BLSTMs do not achieve the same performance as TDNN or TDNN-LSTM models. Note that the number of parameters was (coarsly) optimized for each single acoustic model such that reducing or increasing the number of parameters did not improve the results.

The differences between word and Morfessor-based subword models follow a common pattern, where the subword models outperform the corresponding word-based models. The only exception to this are the *n*-gram models for Swedish which are slightly worse or equal to the word-based result. For RNN-based language models, the subword units show more significant improvements, and the difference between the best subword and word model is more substantial than for the *n*-gram models.

For Finnish and Arabic it has been shown before that subword models can outperform word-based models. Our results validate that this is still the case in a modern WFST-based HMM/DNN speech recognizer, especially when RNN-based language models are used. For Swedish and English this is, as far as we are aware, the first time that Morfessor-subword models outperform a word-based model on a large vocabulary system trained on a large language modeling corpus. While the single best result for English, 17.3% WER, is worse than the best result for a phoneme-based system (16.6%, see Table 6), the system combination of various grapheme-based models improve the results significantly. This will be discussed in Section 4.

Table 8
Word error rates for different English categories using the RNN ‘deep’ language model.

		category					
	Total	advice	childrens	comedy	documentary	drama	events
% of total		14.3	11.3	23.1	35.3	6.1	10.0
char	18.1	14.5	16.9	25.7	14.5	14.9	22.1
morf 0.1	17.4	14.5	**16.0**	**24.4**	**13.8**	**14.6**	**21.5**
word	17.7	**14.4**	16.1	25.0	14.1	14.8	22.0

Table 9
Amount of data used for the different languages.

	Finnish	Arabic	Swedish	English
Speech (training)	1500h	1020h	354h	283h
Speech (evaluation)	5.4h	8.4h	8.7h	5.7h
Text (training tokens)	143M	121M	398M	646M
Unique tokens	4.3M	1.3M	3.5M	0.8M

When comparing the character-based results to the word-based results we see an improvement over word-based models only for Finnish (14.6% vs 13.8%). In Arabic and Swedish the RNN-based character models match the corresponding word-models (16.5% and 2.5% resp.) and for English there is a small degradation (17.7% vs 18.1%). These results are still impressive, as they were made with much smaller RNN models (because of the reduced size of the in- and ouput layers) that took less time to train, especially when the time needed to cluster the vocabulary of a word-based model is taken into account. Another thing to note about the character-based models is that they had the largest relative improvement between *n*-gram and RNN-based models, as well as the most significant relative improvement between “shallow" and "deep" RNN models. This indicates that further refinement of the RNN-based language models might lead to even better results, possibly outperforming word-based models for all languages.

The Morfessor-based subword models did outperform the character models for all languages, indicating that the subword units learned by Morfessor are very suitable for the ASR task. Among the tested lexicon sizes (see Table 2), the number of units optimal for the present tasks varied per language, with the extremes of 10k units for Finnish and 100k units for English. The variation may sound large, but as illustrated by Fig. 2, the actual change in segmentation granularity is not drastic. Similar variation in the best-performing lexicon sizes have been observed also earlier (e.g. Enarvi et al., 2017). Moreover, the differences in word error rates were small, and we expect that for any language a subword lexicon with a size between 10k and 100k will work reasonably well and provide accuracy that is close to the selected ones.

4. System combination between different units

Introducing different sizes of subword units gives a possibility to make a variety of models to evaluate. Not only can these models be evaluated separately, but it is also possible to combine these results using system combination. The most common technique is to use Minimum Bayes Risk (MBR) decoding (Goel and Byrne, 2000; Xu et al., 2010). Instead of optimizing the sentence-error-rate, which is done in ordinary Maximum a Posterori (MAP) decoding, MBR decodes the utterance in such way that minimizes the word error rate directly. This is very powerful in combination with lattice combination, as for each word location the language model likelihood can be averaged between the lattices.

To make system combination work between systems trained on different units, the lattices need to be first converted to the same vocabulary. This can be done with a simple FST-based transformation that is created by first reconstructing all the words in the subword lattice and mapping these subwords to their actual words. Once all lattices are using the same vocabulary, they can be combined into a new lattice that contains all word-paths from the different systems. Afterwards, a standard MBR-decoder is used to get the optimal word sequence.

For all languages evaluated in this paper, we have run system combination experiments in two dimensions. First, we run system combination separately for each language model unit, combining over the three different acoustic models we trained earlier. This shows the improvement for the typical setup where the system combination is used over multiple acoustic models. The second experiment we did was to combine the RNN-deep models for all different units: characters, Morfessor subwords, and words. In addition to doing this for a single acoustic model, we also combined all these units across the two best and all acoustic models.

Table 10 (a) shows the system combination results for the Finnish models. The rightmost column, which shows the results of combinations across acoustic models, indicates that for any unit a significant gain can be made by combining the different AMs. The sharpest decrease in WER is obtained for the *morf 0.01* models, for which the best non-combined result 14.1% is reduced to 12.7% with system combination.

The last rows of Table 10(a) shows the results of system combination over different units. The first result, which combines the 5 TDNN+BLSTM systems improves the best individual result of 13.1% to 12.8%. A combination over multiple acoustic models enhances the result even further to 11.4%. Note that the TDNN model performs best when using RNNLMs, in contrast to TDNN +BLSTM for *n*-gram-based models (Table 3(a)).

For Arabic, the results in Table 10(b) show a similar improvement through system combination. The best cross-AM combination improved the best individual result of 15.6% to 14.9% and combination with multiple units gives the best result of 14.6%. These results are in line with the improvements we obtained in Smit et al. (2017a) where the combination of all our systems improved the best individual result of 15.9% to 14.8%. The main reason for obtaining better results than Smit et al. (2017a) is the use of better (deeper) RNNLMs.

For Swedish, the best WER for an individual system was already very low 2.2%. However, the system combination over different acoustic models reduced this to 2.1% and the combination over both acoustic models and language units to 2.0% WER.

Lastly, the English results are more extensive, as we combined the systems also with the phoneme-based acoustic word model. The best individual result with the phoneme-based model is 16.6% and with the grapheme-based model 17.3% WER. Already when three different grapheme-based acoustic models are combined, the error rate matches or surpasses that of the single phoneme-based system in for all units, except for the character-based model. Combining all systems gives the best WER of

Table 10
Each of the rows ”char, morf, word” contain WER results of one LM trained as the RNN ‘deep’ LM introduced in the chapter III. The four columns are the three AMs and a combination of the corresponding systems (AM comb). The number in parenthesis is the number of systems combined. The LM/AM Comb rows are system combinations of all these LM units, as well as a selection of different AMs. The colored background indicates which of the AMs (columns) are combined.

<table>
<tr><th>Unit</th><th>TDNN</th><th>TDNN+LSTM</th><th>TDNN+BLSTM</th><th>AM comb</th></tr>
<tr><td>char</td><td>13.4</td><td>13.6</td><td>13.8</td><td>12.3 (3)</td></tr>
<tr><td>morf 0.001</td><td>12.8</td><td>13.3</td><td>13.1</td><td>11.9 (3)</td></tr>
<tr><td>morf 0.01</td><td>14.1</td><td>14.5</td><td>14.2</td><td>12.7 (3)</td></tr>
<tr><td>morf 0.1</td><td>13.8</td><td>14.3</td><td>14.0</td><td>12.4 (3)</td></tr>
<tr><td>word</td><td>14.6</td><td>15.0</td><td>14.6</td><td>13.3 (3)</td></tr>
<tr><td></td><td></td><td></td><td>12.8 (5)</td><td></td></tr>
<tr><td>LM/AM</td><td>12.4 (5)</td><td></td><td></td><td></td></tr>
<tr><td>Comb</td><td></td><td colspan="2">11.8 (10)</td><td></td></tr>
<tr><td></td><td colspan="3">11.4 (15)</td><td></td></tr>
</table>

(a) Finnish

<table>
<tr><th>Unit</th><th>TDNN</th><th>TDNN+LSTM</th><th>TDNN+BLSTM</th><th>AM comb</th></tr>
<tr><td>char</td><td>18.9</td><td>17.8</td><td>16.5</td><td>15.9 (3)</td></tr>
<tr><td>morf 0.001</td><td>17.7</td><td>16.7</td><td>15.7</td><td>14.9 (3)</td></tr>
<tr><td>morf 0.01</td><td>17.7</td><td>16.6</td><td>15.6</td><td>14.9 (3)</td></tr>
<tr><td>morf 0.1</td><td>18.4</td><td>17.1</td><td>16.2</td><td>15.4 (3)</td></tr>
<tr><td>word</td><td>18.7</td><td>17.6</td><td>16.5</td><td>15.7 (3)</td></tr>
<tr><td></td><td></td><td></td><td>15.5 (5)</td><td></td></tr>
<tr><td>LM/AM
Comb</td><td></td><td colspan="2">14.7 (10)</td><td></td></tr>
<tr><td></td><td colspan="3">14.6 (15)</td><td></td></tr>
</table>

(b) Arabic

<table>
<tr><th>Unit</th><th>TDNN</th><th>TDNN+LSTM</th><th>TDNN+BLSTM</th><th>AM comb</th></tr>
<tr><td>char</td><td>2.4</td><td>2.5</td><td>3.2</td><td>2.2 (3)</td></tr>
<tr><td>morf 0.001</td><td>2.2</td><td>2.3</td><td>3.1</td><td>2.1 (3)</td></tr>
<tr><td>morf 0.01</td><td>2.4</td><td>2.5</td><td>3.1</td><td>2.2 (3)</td></tr>
<tr><td>morf 0.1</td><td>2.3</td><td>2.4</td><td>2.9</td><td>2.1 (3)</td></tr>
<tr><td>word</td><td>2.4</td><td>2.5</td><td>3.1</td><td>2.2 (3)</td></tr>
<tr><td></td><td></td><td>2.2 (5)</td><td></td><td></td></tr>
<tr><td>LM/AM
Comb</td><td colspan="2">2.0 (10)</td><td></td><td></td></tr>
<tr><td></td><td colspan="3">2.0 (15)</td><td></td></tr>
</table>

(c) Swedish

<table>
<tr><th>Unit</th><th>Phone TDNN</th><th>TDNN</th><th>TDNN+LSTM</th><th>TDNN+BLSTM</th><th>AM comb</th></tr>
<tr><td>char</td><td></td><td>18.1</td><td>18.7</td><td>21.2</td><td>16.9 (3)</td></tr>
<tr><td>morf 0.001</td><td></td><td>17.8</td><td>18.3</td><td>21.0</td><td>16.6 (3)</td></tr>
<tr><td>morf 0.01</td><td></td><td>17.5</td><td>17.9</td><td>20.7</td><td>16.4 (3)</td></tr>
<tr><td>morf 0.1</td><td></td><td>17.3</td><td>17.8</td><td>20.6</td><td>16.2 (3)</td></tr>
<tr><td>word</td><td>16.6</td><td>17.7</td><td>18.2</td><td>20.9</td><td>16.5 (3)</td></tr>
<tr><td></td><td></td><td>17.2 (5)</td><td></td><td></td><td></td></tr>
<tr><td>LM/AM
Comb</td><td></td><td colspan="2">16.3 (10)</td><td></td><td></td></tr>
<tr><td></td><td></td><td colspan="3">16.2 (15)</td><td></td></tr>
<tr><td></td><td colspan="4">15.9 (16)</td><td></td></tr>
</table>

(d) English. The AM comb column combination is only over grapheme models.

15.9%, a 8.0% relative improvement over the best single grapheme-based system and a 4.2% relative improvement over the phoneme-based system. Note that this comparison only indicates the power of combining diverse models, in larger extent across acoustic models and in smaller extent across different units. It also would be possible—and probably effective—to do system combinations across different acoustic models for phoneme-based models. However, for a phoneme-based system it is much more difficult to train models with different kind of units to be used for system combinations From the perspective of low-resource languages that do not have manually created lexicon resources, it is encouraging that even for a language with as irregular pronunciations as English, a normal phoneme-based system can be matched by multiple grapheme-based systems.

Table 11
Comparison of different type of subword language models with 10% language modeling data. The last two columns show the deep RNNLM result for the full data and its relative difference to the deep RNNLM result with 10% data.

		RNNLM		100%	rel. diff
Segmentation	*n*-gram	shallow	deep	deep	to 10%
char	18.5	15.9	15.4	13.8	−10.4%
morf 0.001	17.4	15.4	15.4	13.1	−14.9%
morf 0.01	17.5	16.0	15.7	14.2	−9.5%
morf 0.1	17.6	15.3	15.1	14.0	−7.3%
word	19.6	18.0	17.5	14.6	−16.6%

(a) Finnish

		RNNLM		100%	rel. diff
Segmentation	*n*-gram	shallow	deep	deep	to 10%
char	19.3	17.7	17.5	16.5	−5.7%
morf 0.001	18.1	17.1	17.0	15.7	−7.7%
morf 0.01	18.1	17.1	17.0	15.6	−8.2%
morf 0.1	18.5	17.2	17.0	16.2	−4.7%
word	19.4	19.4	18.3	16.5	−9.8%

(b) Arabic

		RNNLM		100%	rel. diff
Segmentation	*n*-gram	shallow	deep	deep	to 10%
char	6.0	4.2	3.7	2.5	−32.4%
morf 0.001	5.3	3.4	3.1	2.3	−25.8%
morf 0.01	5.3	3.8	3.5	2.5	−28.6%
morf 0.1	5.4	3.6	3.4	2.4	−29.4%
word	6.0	4.7	4.4	2.5	−43.2%

(c) Swedish

		RNNLM		100%	rel. diff
Segmentation	*n*-gram	shallow	deep	deep	to 10%
char	21.5	19.8	19.1	18.1	−5.2%
morf 0.001	20.6	18.4	18.2	17.8	−2.2%
morf 0.01	20.4	19.0	18.7	17.5	−6.4%
morf 0.1	20.4	18.6	18.5	17.3	−6.5%
word	22.0	19.3	18.8	17.7	−5.9%

(d) English

To further optimize the recognition accuracy we could also try system combination over different phoneme-based acoustic models as we did for grapheme-based models. However, with the current experiments we can already conclude that our grapheme-based systems are able to capture the different pronunciations of graphemes with surprising accuracy. To our knowledge, it has not been shown before that the difference between grapheme and phoneme-based systems can be this small for the English language.

5. Under-resourced scenario

As explained in Section 2.4, one of the strengths of subword models, including character models, is that the language model training data contains more examples of each unit and therefore can be more effectively used in language model training compared to the word models. To demonstrate this, we trained all systems with only 10% of the available language modeling data and compared the impact on the performance for different language modeling units. This effectively simulates an under-resourced scenario where less language modeling data is available.

In this study we reduced only the amount of language modeling data and kept the acoustic models identical to the previous experiments. Naturally, the acoustic model will also learn some (grapheme-based) language patterns through the sequence-

Table 12
Comparison to previously published results.

Language	Ours	Prev	Reference	Differences to ours
Finnish	11.4	28.4	Kurimo et al. (2017)	GMM-HMM acoustic model. Similar LM with vocabulary adaptation
Arabic	14.6	16.2	Manohar et al. (2017)	Uses sMBR in acoustic model. Word-based 4-gram LM.
Swedish	2.0	15.97	Kaldi Sprakbanken recipe	Cross-entropy acoustic model. Word-based 4-gram LM.
English	15.9	17.9	Wang et al. (2018)	Also mix of grapheme + phoneme. Word-based RNN LM.

based training criterion, but we do not expect this to affect the experiment: The language modeling data does not contain the acoustic model training utterances, and the number of word types and tokens in the acoustic modeling data is limited. In the Finnish experiment, where we have the largest acoustic modeling training set, there are still only 9M tokens and 400k unique words present in the acoustic training data. Even if this data were counted into the amount of language modeling data, the total amount of data used would still be less than 20% of the original text data.

For Finnish (Table 11(a)), the results are between 7 and 17% worse than with the full language modeling data. As hypothesized, the results of the subword models degrade less than those of the word-based models. The character model is now outperforming both the word model as well as two of the Morfessor subword models.

Table 11 (b) shows a similar pattern for the Arabic data set. The character model outperforms the word-based model even for the *n*-gram model.

In Swedish, as shown in Table 11(c), the relative degradations seem larger than for the other tasks. This may be caused by the nature of errors when the WER is much closer to zero, compared to the other tasks. In the under-resourced case, the character model outperforms the word-based model and comes close to the performance of the Morfessor-based subword models.

Unlike in the other languages, for English the performance of the subword models stays close to the word models both in the 10% and in the full 100% data. On the other hand, the amount of language modeling data for English was the largest (see Table 9), with the 10% system still containing 65M training tokens.

For all languages, there is a clear degradation in results when the amount of language modeling data is reduced. As expected, the subword models are more robustly handling the data sparsity in all languages. The only exception to this, English, can be explained by the fact that the original vocabulary is much smaller and the language modeling data much more substantial than for the other languages. When comparing the character-based models to the Morfessor subword models, only in Finnish their performance matches to two out of three Morfessor models, with the best Morfessor model still outperforming the character based model (15.1% vs 15.4%). In other languages, the character models still stay behind.

6. Conclusion

We set out to implement and evaluate the use of subword units in state-of-the-art speech recognition, including advanced neural network based acoustic and language models. To do this, we have evaluated word and subword systems for four distinct languages from different language families.

For all four datasets, we improved over previously published results. Table 12 compares our results to the best results published elsewhere. Note that for Swedish, we could not find any paper referencing this dataset, but instead we have compared to the generally available ‘sprakbanken’ recipe of the Kaldi toolkit. For Finnish, we have omitted results that were published in our own work (Smit et al., 2017c; 2017b) using mainly the same techniques. For Arabic, the ‘previous’ best published result is from MGB-3 challenge (Ali et al., 2017; the MGB-2 dataset), in which our submission (Smit et al., 2017a) using the techinques presented here reached the top scores.

Our primary evaluation shows that models based on subwords derived by Morfessor consistently outperform word-based models on all tested languages. The optimal size of the subword lexicon varied across languages, from less than 10,000 in Finnish to a bit over 100,000 in English, but in no situation did subword models perform worse than the same model with word units. Although this effect is already present when using *n*-gram language models, it is even stronger for RNN-based language models, which have a better capability of capturing longer contexts.

With accurate RNN language models, using single characters as subword units also yields surprisingly good results. A character-based model outperformed the word-based model for Finnish, produced a similar performance for Arabic and Swedish, and underperformed the word-based model only for English.

As using different subword units provides with a variety of different models, we combined these systems using MBR-based system combination. Although it was already effective to combine acoustic models from three different architectures, using models with different language modeling units was also a success, with combination systems reducing the WER of the single best system with over 10% relative.

Although it was not the original intent of this paper, we did obtain interesting results regarding the use of grapheme models for English. Not only was a simple direct word-based comparison between grapheme and phoneme models only showing a 6% edge for phoneme-based models, by doing system combination the results of the grapheme-based system could match and surpass the phoneme-based result, without having the need for a hand-crafted pronuciation lexicon. Furthermore, a combination with all the grapheme and phoneme-based systems that we trained resulted in a 15.9% WER for the MGB dev17b test set, which is by far the best result published on this data set.

Lastly, the evaluation with smaller language modeling data sets showed another strength of the subwords, where the degradation was significantly smaller than that of the word units. This confirms the hypothesis that using subword units reduces data sparsity and increases model robustness.

Acknowledgements

This work was supported by Svenska folkskolans vänner r.f. via the DigiTala project, Business Finland's Challenge Finland project TELLme, Kone foundation, EU's Horizon 2020 research and innovation programme via the project MeMAD (GA 780069) and the project FoTran (GA 771113). Computational resources were provided by the Aalto Science-IT project and the CSC – IT Center for Science, Finland.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ali, A., Bell, P., Glass, J., Messaoui, Y., Mubarak, H., Renals, S., Zhang, Y., 2016. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. SLT 2016 – IEEE Spoken Language Technology Workshop, pp. 279–284. https://doi.org/10.1109/SLT.2016.7846277.

Ali, A., Mubarak, H., Vogel, S., 2014. Advances in dialectal Arabic speech recognition: a study using twitter to improve Egyptian ASR. IWSLT 2014 – International Workshop on Spoken Language Translation.

Ali, A., Vogel, S., Renals, S., 2017. Speech recognition challenge in the wild: Arabic MGB-3. 2017 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), Okinawa, pp. 316–322.

Arisoy, E., Can, D., Parlak, S., Sak, H., Saraclar, M., 2009. Turkish broadcast news transcription and retrieval. IEEE Trans Audio Speech Lang Process 17 (5), 874–883. https://doi.org/10.1109/TASL.2008.2012313.

Bisani, M., Ney, H., 2005. Open vocabulary speech recognition with flat hybrid models. INTERSPEECH 2005 – Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, pp. 725–728.

Botros, R., Irie, K., Sundermeyer, M., Ney, H., 2015. On efficient training of word classes and their application to recurrent neural network language models. INTERSPEECH (a), pp. 1443–1447.

Chan, W., Jaitly, N., Le, Q., Vinyals, O., 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. ICASSP 2016 – IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4960–4964. https://doi.org/10.1109/ICASSP.2016.7472621.

Cheng, G., Peddinti, V., Povey, D., Manohar, V., Khudanpur, S., Yan, Y., 2017. An exploration of dropout with lstms. pp. 1586–1590. https://doi.org/10.21437/Interspeech.2017-129.

Choueiter, G., Povey, D., Chen, S.F., Zweig, G., 2006. Morpheme-based language modeling for Arabic LVCSR. ICASSP 2006 – IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1053–1056. https://doi.org/10.1109/ICASSP.2006.1660205.

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., Stolcke, A., 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. ACM Trans. Speech Lang. Process. 5 (1), 3:1–3:29. https://doi.org/10.1145/1322391.1322394.

Creutz, M., Lagus, K., 2002. Unsupervised discovery of morphemes. In: Proceedings of the ACL 2002 Workshop on Morphological and Phonological Learning. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 21–30. https://doi.org/10.3115/1118647.1118650.

Creutz, M., Lagus, K., 2007. Unsupervised models for morpheme segmentation and morphology learning. ACM Trans. Speech Lang. Process. 4 (1).

CSC - IT Center for Science, 1998. The Helsinki Korp Version of the Finnish Text Collection. URL: http://urn.fi/urn:nbn:fi:lb-2016050207

Enarvi, S., Kurimo, M., 2016. TheanoLM – an extensible toolkit for neural network language modeling. INTERSPEECH (b), pp. 3052–3056. https://doi.org/10.21437/Interspeech.2016-618.

Enarvi, S., Smit, P., Virpioja, S., Kurimo, M., 2017. Automatic speech recognition with very large conversational Finnish and Estonian vocabularies. IEEE/ACM Trans Audio Speech Lang Process 25 (11), 2085–2097. https://doi.org/10.1109/TASLP.2017.2743344.

Gage, P., 1994. A new algorithm for data compression. The C Users Journal 12 (2), 23–38.

Goel, V., Byrne, W.J., 2000. Minimum Bayes-risk automatic speech recognition. Computer Speech & Language 14 (2), 115–135. https://doi.org/10.1006/csla.2000.0138.

Graves, A., Jaitly, N., 2014. Towards end-to-end speech recognition with recurrent neural networks. In: Xing, E.P., Jebara, T. (Eds.), Proceedings of the 31st International Conference on Machine Learning. PMLR, Bejing, China, pp. 1764–1772.

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J., 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. Computer Speech & Language 20 (4), 515–541. https://doi.org/10.1016/j.csl.2005.07.002.

Hirsimäki, T., Pylkkönen, J., Kurimo, M., 2009. Importance of high-order n-gram models in morph-based speech recognition. IEEE Trans Audio Speech Lang Process 17 (4), 724–732. https://doi.org/10.1109/TASL.2008.2012323.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput 9 (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Iskra, D.J., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F., Kiessling, A., 2002. SPEECON-Speech databases for consumer devices: Database specification and validation. LREC.

Kirchhoff, K., Vergyri, D., Bilmes, J., Duh, K., Stolcke, A., 2006. Morphology-based language modeling for conversational Arabic speech recognition. Computer Speech & Language 20 (4), 589–608. https://doi.org/10.1016/j.csl.2005.10.001.

Kneser, R., Ney, H., 1993. Forming word classes by statistical clustering for statistical language modelling. In: Köhler, R., Rieger, B.B. (Eds.), Contributions to Quantitative Linguistics. Kluwer Academic Publishers, Dordrecht, the Netherlands, pp. 221–226. https://doi.org/10.1007/978-94-011-1769-2_15.

Kuo, H.-K., Arısoy, E., Emami, A., Vozila, P., 2012. Large scale hierarchical neural network language models. INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association, pp. 1672–1675.Portland, OR, USA.

Kurimo, M., Enarvi, S., Tilk, O., Varjokallio, M., Mansikkaniemi, A., Alumäe, T., 2017. Modeling under-resourced languages for speech recognition. Lang Resour Eval 51 (4), 961–987.

Manohar, V., Povey, D., Khudanpur, S., 2017. Jhu kaldi system for Arabic mgb-3 asr challenge using diarization, audio-transcript alignment and transfer learning. ASRU, pp. 346–352.

Mansikkaniemi, A., Smit, P., Kurimo, M., 2017. Automatic construction of the Finnish Parliament Speech Corpus. Proc. Interspeech 2017, pp. 3762–3766.

Mikolov, T., Karafit, M., Burget, L., Cernocký, J., Khudanpur, S., 2010. Recurrent neural network based language model. INTERSPEECH 2010 – 11th Annual Conference of the International Speech Communication Association, Makuhari, Japan, pp. 1045–1048.

Mirjam, K., Sebastian Stuker, T.S., 2003. Grapheme based speech recognition. INTERSPEECH 2003 – Eurospeech, 8th European Conference on Speech Communication and Technology.Geneva, Switzerland.

Mohri, M., Pereira, F., Riley, M., 2008. Speech recognition with weighted finite-state transducers. Springer Handbook of Speech Processing. Springer, pp. 559–584.

Morin, F., Bengio, Y., 2005. Hierarchical probabilistic neural network language model. Aistats, 5. Citeseer, pp. 246–252.

Mousa, A.E.-D., Kuo, H.-K.J., Mangu, L., Soltau, H., 2013. Morpheme-based feature-rich language models using deep neural networks for LVCSR of Egyptian Arabic. ICASSP 2013 – IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8435–8439. https://doi.org/10.1109/ICASSP.2013.6639311.

Peddinti, V., Povey, D., Khudanpur, S., 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. INTERSPEECH (a), pp. 3214–3218.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The Kaldi speech recognition toolkit. ASRU 2011 – IEEE Workshop on Automatic Speech Recognition & Understanding.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., Khudanpur, S., 2016. Purely sequence-trained neural networks for ASR based on lattice-free MMI. INTERSPEECH (b), pp. 2751–2755. https://doi.org/10.21437/Interspeech.2016-595.

Prabhavalkar, R., Rao, K., Sainath, T.N., Li, B., Johnson, L., Jaitly, N., 2017. A comparison of sequence-to-sequence models for speech recognition. INTERSPEECH (c), pp. 939–943. https://doi.org/10.21437/Interspeech.2017-233.

Rao, K., Sak, H., 2017. Multi-accent speech recognition with hierarchical grapheme based models. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4815–4819. https://doi.org/10.1109/ICASSP.2017.7953071.

Rissanen, J., 1978. Modeling by shortest data description. Automatica 14 (5), 465–471.

Rosti, A., Rämö, A., Saarelainen, T., Yli-Hietanen, J., 1998. SpeechDat Finnish Database for the fixed telephone network. Technical Report. Tampere University of Technology.

Sainath, T.N., Prabhavalkar, R., Kumar, S., Lee, S., Kannan, A., Rybach, D., Schogol, V., Nguyen, P., Li, B., Wu, Y., Chen, Z., Chiu, C., 2018. No need for a lexicon? Evaluating the value of the pronunciation lexica in end-to-end models. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5859–5863. https://doi.org/10.1109/ICASSP.2018.8462380.

Sennrich, R., Haddow, B., Birch, A., 2015. Neural machine translation of rare words with subword units. ACL16.ArXiv: 1508.07909.

Siivola, V., Hirsimäki, T., Virpioja, S., 2007. On growing and pruning Kneser-Ney smoothed N-gram models. IEEE Transactions on Audio, Speech & Language Processing 15 (5), 1617–1624.

Smit, P., Gangireddy, S.R., Enarvi, S., Virpioja, S., Kurimo, M., 2017. Aalto system for the 2017 Arabic multi-genre broadcast challenge. ASRU.

Smit, P., Gangireddy, S.R., Enarvi, S., Virpioja, S., Kurimo, M., 2017. Character-based units for unlimited vocabulary continuous speech recognition. ASRU.

Smit, P., Leinonen, J., Jokinen, K., Kurimo, M., 2016. Automatic speech recognition for northern sámi with comparison to other uralic languages. In: Pirinen, T.A., Simon, E., Tyers, F.M., Vincze, V. (Eds.), Second International Workshop on Computational Linguistics for Uralic Languages. University of Szeged, Szeged, Hungary, p. 12.

Smit, P., Virpioja, S., Kurimo, M., 2017. Improved subword modeling for WFST-based speech recognition. INTERSPEECH (c).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15 (1), 1929–1958.

Srivastava, R.K., Greff, K., Schmidhuber, J., 2015. Training very deep networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 2377–2385.

Tarján, B., Fegyó, T., Mihajlik, P., 2014. A bilingual study on the prediction of morph-based improvement. SLTU, pp. 131–138.

Varjokallio, M., Kurimo, M., Virpioja, S., 2013. Learning a Subword Vocabulary Based on Unigram Likelihood. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc, Czech Republic, pp. 7–12.

Virpioja, S., Kohonen, O., Lagus, K., 2011. Evaluating the Effect of Word Frequencies in a Probabilistic Generative Model of Morphology. In: Pedersen, B.S., Nešpore, G., Skadiņa, I. (Eds.), Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011). NEALT Proceedings Series. 11, Northern European Association for Language Technology, Riga, Latvia, pp. 230–237.

Virpioja, S., Smit, P., Grönroos, S.-A., Kurimo, M., 2013. Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. Report. Department of Signal Processing and Acoustics, Aalto University. Helsinki, Finland.

Wang, Y., Chen, X., Gales, M., Ragni, A., Wong, J., 2018. Phonetic and graphemic systems for multi-genre broadcast transcription. ICASSP 2018 – IEEE International Conference on Acoustics, Speech and Signal Processing.

Wong, J.H.M., Gales, M.J.F., 2017. Multi-task ensembles with teacher-student training. ASRU, pp. 84–90. https://doi.org/10.1109/ASRU.2017.8268920.

Xu, H., Povey, D., Mangu, L., Zhu, J., 2010. An improved consensus-like method for minimum Bayes risk decoding and lattice combination. ICASSP 2010 – IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4938–4941. https://doi.org/10.1109/ICASSP.2010.5495100.