Effects of Language Relatedness for Cross-lingual Transfer Learning in Character-Based Language Models

Mittul Singh[*], **Peter Smit**[*‡], **Sami Virpioja**[†], **Mikko Kurimo**[*]

[*]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[†]Department of Digital Humanities, Helsinki University, Helsinki, Finland
[‡]Inscripta, Helsinki, Finland

`firstname.lastname@{aalto,helsinki}.fi`

Abstract

Character-based Neural Network Language Models (NNLM) have the advantage of smaller vocabulary and thus faster training times in comparison to NNLMs based on multi-character units. However, in low-resource scenarios, both the character and multi-character NNLMs suffer from data sparsity. In such scenarios, cross-lingual transfer has improved multi-character NNLM performance by allowing information transfer from a *source* to the *target* language. In the same vein, we propose to use cross-lingual transfer for character NNLMs applied to low-resource Automatic Speech Recognition (ASR). However, applying cross-lingual transfer to character NNLMs is not as straightforward. We observe that relatedness of the source language plays an important role in cross-lingual pretraining of character NNLMs. We evaluate this aspect on ASR tasks for two target languages: Finnish (with English and Estonian as source) and Swedish (with Danish, Norwegian, and English as source). Prior work has observed no difference between using the related or unrelated language for multi-character NNLMs. We, however, show that for character-based NNLMs, only pretraining with a related language improves the ASR performance, and using an unrelated language may deteriorate it. We also observe that the benefits are larger when there is much lesser target data than source data.

Keywords: Cross-lingual transfer, Character language models, Low-resource ASR

1. Introduction

Multilingual training of language models has successfully leveraged datasets from other languages to improve Neural Network Language Modeling (NNLM) performance in low-resource scenarios (Kim et al., 2019; Conneau and Lample, 2019; Conneau et al., 2019; Aharoni et al., 2019). One such method for training NNLM is the multi-task-based approach, where multiple language corpora train the model simultaneously (Aharoni et al., 2019). Another approach is cross-lingual pretraining, where the NNLM is trained on a set of *source* languages followed by fine-tuning on the *target* language (Kim et al., 2019; Conneau and Lample, 2019; Conneau et al., 2019). The second approach, explored in this work, is favorable when re-training with the large source data is time-consuming as an existing trained source model's weights can be transferred to the target model and then fine-tuned on the smaller target data. Cross-lingually pretrained NNLMs have utilized multi-character units to construct large shared vocabulary to allow the positive transfer of information from source to target. Instead of multi-character units, we explore a single character as a modeling unit for applying cross-lingual pretraining. This choice has the advantage of reducing the vocabulary size by several orders of magnitude and providing a larger intersection of vocabulary terms than multi-character units. In this paper, we apply cross-lingual pretraining to character NNLMs. However, this off-the-shelf application is not trivial. For multi-character based NNLMs, cross-lingual pretraining works by sharing information across various source languages independent of relatedness to the target language in terms of closeness in the language family tree[1]. In contrast, for character-based NNLMs, a source

language in the same family subtree as the target (related) affects the downstream performance positively than from an unrelated source language.

We experiment with available Finnish and Swedish Automatic Speech Recognition (ASR) systems in a simulated low-resource ASR scenario by limiting the language modeling resources. We apply pretraining with two source languages (Estonian and English) for Finnish ASR and three source languages (Danish, English, and Norwegian) for Swedish ASR. In our experiments, we observe perplexity and ASR performance improvements when pretraining NNLMs with related languages (i.e. Estonian for Finnish and Danish and Norwegian for Swedish), whereas pretraining NNLMs on English performs adversely.

We also study the impact on cross-lingual transfer due to the target data size and number of source model layers transferred. Relatively, smaller amounts of target language data than the source language data leads to more considerable ASR performance improvements. Moreover, we find that pretrained NNLMs perform best when we transfer only the parameters of the lowest layer of the source model.

2. Related Work

In our work, we follow the cross-lingual pretraining scheme utilizing a shared vocabulary as proposed by Zhuang et al. (Zhuang et al., 2017), where they transfer all the hidden layers except the final layer from the source model to the target model. For NNLMs, such an application does not obtain the best results. In sections 6. and 7., we present results to support this observation.

Concurrently, Lample and Conneau (Conneau and Lample, 2019) have also shown that cross-lingual pretraining can improve the performance of language models on intrinsic measures like perplexity. They train a multi-character

---

[1]`https://en.wikipedia.org/wiki/Language_family`

Language	Vocabulary	Train	Dev
Finnish ASR			
English (En)	232K	116M	107K
Estonian (Et)	1.7M	97M	33K
Finnish (Fi)	1.1M	17M	130K
Swedish ASR			
Danish (Da)	2.7M	365M	222K
English (En)	466K	366M	107K
Norwegian (No)	2.4M	381M	194K
Swedish (Sv)	936K	45M	158K

Thousands (K), Millions (M)

Table 1: The table reports the word vocabulary, training set (Train) and development set (Dev) sizes of the languages used in the experiments.

transformer-based language model with a masked language model training procedure for cross-lingual pretraining. In their model, multi-character units from both the source and target languages are combined to form one large vocabulary. This large shared vocabulary leads to a large output layer, which can be inefficient to train. The layer size can be reduced by shortlists and class-based models (Goodman, 2001; Le et al., 2011), or approximated by applying a hierarchical softmax (Morin and Bengio, 2005). Instead, we choose characters as the basic unit of modeling, which provides a more natural way of reducing the vocabulary size. Simultaneously, this choice supports the cross-lingual information transfer by providing a larger intersection of vocabulary terms than multi-character units.

For cross-lingual pretraining, language relatedness remains an unexplored factor, which becomes the focus of our work. Prior work has applied cross-lingual transfer by using several unrelated languages as a source. Using related language can be crucial in low-resource scenarios as we discover in Section 6. and 7. In our work, we limit cross-lingual transfer from one source language allowing a simpler setup for better analysis, in future, we would like to explore the impact of relatedness when the number of source languages is increased dramatically.

3. Datasets

We create two setups to evaluate cross-lingual pretraining for NNLMs. In the first setup, English (En) and Estonian (Et) are the high-resource sources of language modeling corpora, and Finnish (Fi) is the low-resource target language. In the second setup, Danish (Da), English, and Norwegian (No) are the high-resource source languages, and Swedish (Sv) is the low-resource target language.

Estonian and Finnish are contained in the Finnic language subtree, and Danish, Norwegian, and Swedish belong to the North Germanic language subtree. Thus, these source-target set of languages are considered as related languages. For both Finnish and Swedish, English, being part of the West Germanic language subtree, is considered as a more unrelated language. We also chose English as it has a large intersection for the character set, but is less mutually intelligible in comparison.

The English text is obtained from the training data of 2015 MGB Challenge (Bell et al., 2015) consists of BBC news transcripts. The Estonian corpus consists of web crawl text
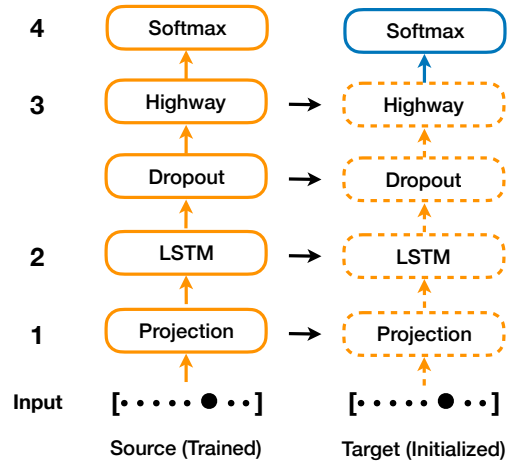

Figure 1: The figure displays the source and target NNLMs with the hidden layers used in our experiments. In cross-lingual pretraining, the source-language-trained hidden layers initialize parts (dotted lines) of the target-language network shown by the arrows. In contrast, the rest is randomly initialized (bold lines in the target network).

and spontaneous conversational transcripts from Meister et al. (2012) and has been used by Enarvi et al. (2017). The Finnish corpus is from Finnish Text Collection containing text from newspaper, books and novels (CSC - IT Center for Science, 1998) and has been used by Smit et al. (2017). The Swedish, Danish, and Norwegian corpora, containing newspaper articles, are downloaded from Språkbanken corpus[2] and have been used by Smit et al. (2018). For Finnish as the target, more data for English was available than for Estonian, so we extract only a portion of English dataset to allow for a similar average of words per line for both datasets. We list the corpora statistics for the various languages used in our experiments in Table 1.

4. Building Language Models

We train character NNLMs for our experiments and mark both the left and right ends of characters except when at the beginning or the end of a word (e.g., model = m+ +o+ +d+ +e+ +l) to achieve best results (Smit et al., 2017). With this marking scheme, we can differentiate the characters from a word into beginning (B), middle (M), end (E) and singleton units. This notation becomes relevant in Section 6., where analyze the differences in perplexity per word position.

We build Recurrent Neural Network Language Models (RNNLM) with a projection layer (200 neurons), an LSTM layer (1000 neurons), a highway layer (1000 neurons) and a softmax output layer (displayed in Figure 1). In our experiments, both the source- and target-language neural networks have the same architecture. We train the RNNLMs using TheanoLM (Enarvi and Kurimo, 2016), applying the adaptive gradient (Adagrad) algorithm to update the model parameters after processing a mini-batch of training examples. The mini-batch size for models was 64, with a sequence length of 100. We used an initial learning rate of 0.1 in all the experiments and a dropout of 0.2 was used to regularize the parameter learning.

---

[2]https://www.nb.no/sprakbanken

Finnish Test Set Perplexity				
Fi$_0$ (baseline)	3788			
l	4	3	2	1
En→Fi	4195	4617	5458	4211
Et→Fi	3402	3585	3901	**3009**
Swedish Test Set Perplexity				
Sv$_0$ (baseline)	311			
l	4	3	2	1
En→Sv	334	322	337	315
No→Sv	**285**	311	312	287
Da→Sv	291	292	317	291

Table 2: The table reports NNLM's test set perplexity for Finnish and Swedish using different cross-lingual initializations. For Finnish, English and Estonian are used as the source languages for pretraining. For Swedish, we use Danish, English and Norwegian as source languages. The best results in each category are marked in boldface.

5. Exploring Cross-Lingual Pretraining

Cross-lingual pretraining involves first training the neural network on a source language. Then, starting from the input layer, the source network's hidden layers initialize the target-language neural network partially or wholly. In a partial initialization, we initialize the uninitialized layers randomly. This initialization step is followed by training on the target language, also referred to as the fine-tuning step. In both the pretraining and the fine-tuning step, the output-layer vocabulary consists of character units from all the source languages and the target language. The pretraining step transfers coarser-level information from input to higher layers into the target model and during fine-tuning, the target model refines this transferred information to a more fine-grained level.

We study neural network models across three dimensions: **1)** the source language used for pretraining step; **2)** using the number of target-model hidden layers (l) initialized starting from the input layer; and **3)** the amount of target language data. We represent the LM pretrained using the source language y and fine-tuned using target language z as $y \rightarrow z$. We vary l from 1 to 4 for the architecture in Figure 1, which also shows an example for $l = 3$. Here $l = 1$ would refer to just initializing with the projection layer and $l = 4$ would refer to initializing with all the layers. We increase the amount of target data size to match the source data size. Varying these parameters allows us to understand their effect on transfer capacity of cross-lingual pretraining.

6. Perplexity Experiments

Table 2 presents the test set perplexity of Finnish and Swedish LMs. When using related source languages — like Estonian for Finnish, or Danish and Norwegian for Swedish — to pretrain the models, we obtain better perplexity than the baseline and when English (the unrelated source language) is used, which leads to a worse perplexity for all ls. Using related source languages, the pretrained target LMs outperform the baseline results for most values of l but, notably, when initialized with configurations $l = 1, 4$ of the source model. Here, we note that Finnish perplexity values are large due to long words and the domain mismatch be-
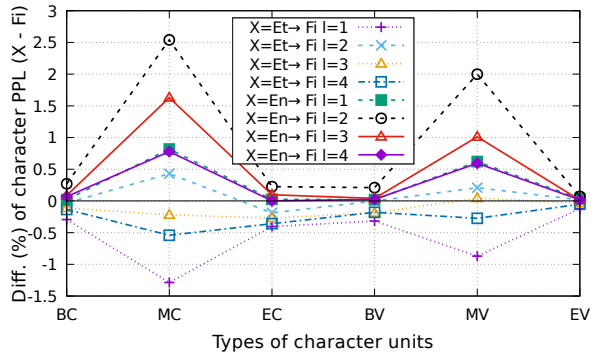

Figure 2: The figure shows the relative differences (%) in character perplexity (PPL) for three different types of character units of different Finnish NNLMs on the test set. These character units exist due to the marking scheme used here: beginning (B), middle (M) and end (E), which can further be classified into consonants (C) and vowels (V).

tween the training (books, newspaper articles and journals) and test (broadcast news) sets.

On characters, similar trends of perplexity improvement for related vs unrelated source language and different values of l are observed. Character perplexity differences for Finnish are presented in Figure 2 for different types of units, i.e., consonant (C) and vowels (V) dependent on their position in words beginning (B), middle (M) and end (E). In Figure 2, most perplexity improvements are obtained for middle consonants (MC) and middle vowels (MV), which are more frequent than other character units. For other character units, small but consistent improvements are obtained by Et→Fi ($l = 1$) LM over other baseline and other LMs. For Swedish, similar improvements to Finnish results are observed for MC and MV, but some dips are seen for Danish-pretrained LMs on end consonants (EC). For brevity, we do not present this result in the paper. Overall, improvements from related-language pretraining impacts the different types of characters, enabled by a large intersection in the source-target character set.

We suspect that pretraining with a related language finds more useful information than with an unrelated one. To investigate this effect, we calculate the cosine similarity between pretrained and baseline LMs' output layer embeddings. We first find an affine transformation to align pretrained LM's embeddings with the baseline's embedding space, and then calculate the average similarity between the two sets. On Finnish, the English-pretrained embeddings have a higher average similarity (0.53) to the baseline embeddings than the Estonian-pretrained embeddings (0.51). On Swedish, similar results are observed with cosine similarity for the English-, Norwegian- and Danish-pretrained embeddings at 0.43, 0.42 and 0.42. They suggest that the related-language pretrained LMs have more conflicting information than the English-pretrained LMs. As they also perform better in terms of perplexity, the related-language pretraining seems to learn information that is complementary to the baseline LM.

7. Speech Recognition Experiments

For training the Finnish acoustic models, we used 1500 hours of Finnish audio from three different sources, namely,

Language	Baseline Architecture			
Fi_0 (baseline)	**16.44**			
l	4	3	2	1
En→Fi	16.70	16.90*	17.34*	16.56
Et→Fi	16.20	16.14*	16.61	**16.01***
Linear interpolations				
En→Fi + Fi_0	16.00*	16.24*	16.34	15.95*
Et→Fi + Fi_0	15.89*	15.87*	16.04*	**15.74***

Table 3: The table reports WER on Finnish ASR task using different cross-lingual initializations for RNNLMs used in rescoring. Here English and Estonian are used as the source languages for pretraining. Asterisks (*) denote statistical significance while comparing against Fi (16.44) using the matched pairs test with $p < 0.05$. The best results in each section are marked in boldface.
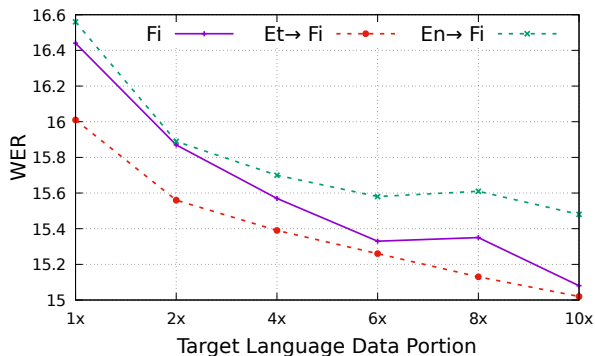

Figure 3: The figures display WERs on Finnish ASR measured when varying the amount of source language data and when varying the amount of target language data.

the Speecon corpus (Iskra et al., 2002), the Speechdat database (Rosti et al., 1998) and the parliament corpus (Mansikkaniemi et al., 2017). For testing, we used a broadcast news dataset from the Finnish national broadcaster (Yle) containing 5 hours of speech and 35k words (Mansikkaniemi et al., 2017). For training Swedish acoustic models, we used 354 hours of audio provided by the Språkbanken corpus. From the original evaluation set, we used a total of 9 hours for development and evaluation.

The acoustic models were trained with the Kaldi toolkit (Povey et al., 2011) with a similar recipe as (Smit et al., 2017). Instead of phonemes, we use grapheme-units, as this allows for a trivial lexicon that maps between the acoustic and language modeling units. We evaluate the ASR performance in terms of Word Error Rates (WER).

For the first-pass, we train a variable-length Kneser-Ney (Kneser and Ney, 1995) n-gram LM using the VariKN toolkit (Siivola et al., 2007). Then, RNNLMs, built in Section 4., are used to rescore the lattices. We also linearly interpolate cross-lingually pretrained NNLMs with target-only NNLM while optimizing the interpolation weight. We test the statistical significance of our results using the Matched Pairs Sentence Segment Word Error Test from NIST Scoring toolkit[3] to compare different systems. Ta-

---

[3]SCTK: http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sctk.htm

Language	Baseline Architecture			
Sv_0 (baseline)	**4.42**			
l	4	3	2	1
En → Sv	4.43	4.46	4.62	4.41
No → Sv	4.18	4.42	4.38	4.17*
Da → Sv	4.24	4.16*	4.39	**4.15***
Linear Interpolations				
En → Sv + Sv_0	4.15*	4.18*	4.20	4.15*
No → Sv + Sv_0	4.02*	4.40	4.35	4.00*
Da → Sv + Sv_0	4.01*	4.04*	4.15*	**3.98***

Table 4: The table reports WER on Swedish ASR task for different configurations of RNNLMs used in rescoring. Here Danish, English and Norwegian are used as the source languages for cross-lingual pretraining. Asterisks (*) denote statistical significance when comparing to Sv (4.41) using the matched pairs test with $p < 0.05$. The best results in each section are marked in boldface.

bles 3 and 4 outline the performance of rescoring with RNNLMs (Section 4.) on a Finnish and a Swedish ASR task. The first row of both these tables displays the performance of target-only trained RNNLMs (baseline). The second part reports the performance of cross-lingually pretrained models (Section 5.) and the third part reports their linear interpolations with target-only baseline models.

Similar to the perplexity results (Section 6.), related source language pretraining improves the ASR performance over the baseline models and the unrelated source language pretraining degrades the performance. On Finnish ASR, English-pretrained RNNLM (En→Fi) lags behind the Estonian-pretrained RNNLM (Et→Fi), which also outperforms Finnish-only models. On Swedish ASR, Danish (Da→Sv) and Norwegian (No→Sv) pretrained models outperform the baseline and English pretrained models (En→Sv). In contrast with perplexity results, lower-layer ($l = 1$) based initialization shows the most benefit over the higher-layer ($l = 2, 3, 4$) initializations for both Finnish and Swedish ASR. We note that quite like perplexity results, ASR performance on Swedish is lower than Finnish as the Swedish task is easier than the Finnish one.

In Figure 3, we observe little performance increase by cross-lingual pretraining when we vary the target data size by increasing it to comparable sizes of source language data. At least for Estonian, increasing Finnish data (target) closes the gap between cross-lingual pretraining and target-only model. The cross-lingual transfer seems to work best with a larger number of resources for the related source language in comparison to the target language.

Furthermore, interpolations of the baseline model with the cross-lingually pretrained models improve over its constituent models. On both Finnish and Swedish ASR, cross-lingual pretraining with English combined with the baseline model can outperform the baseline model, unlike when used individually. This improvement can be attributed to the regularization effect of such an interpolation. Linear interpolations based on other source languages like Estonian, Danish and Norwegian further improve the results consistently across different initialization schemes. We hypothesize that this effect is due to the complementary informa-

tion learned by these related-language models. Overall, the individual systems and the interpolations based on related source languages show a significant and the most substantial improvement in performance.

8. Concluding Remarks

We investigated cross-lingual transfer for character-based neural network language models in a low-resource scenario. Cross-lingual pretraining with related source language significantly improved (3-6% relative) over no pretraining, whereas pretraining with unrelated source language had adverse effects. At a character level, we suspect cross-lingual pretraining works for related languages as they share a large portion of the character set. The large shared vocabulary provides soft alignments between characters in related languages supporting the transfer of relevant information from source to target models. This information transfer is in contrast to multi-character units where the transfer is dependent on shared anchor tokens (like numbers, proper nouns). However, we still lack an empirical understanding of this phenomenon and in our future work, we hope to explore this phenomenon.

Additionally, transferring the lower layer information and having more source data than target data was significant for low-resource ASR. As a followup to our study, we investigate the effects of language relatedness for cross-lingual pretraining in transformer-based language models.

9. Bibliographical References

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bell, P., Gales, M. J., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., McParland, A., Renals, S., Saz, O., Wester, M., et al. (2015). The mgb challenge: Evaluating multi-genre broadcast media recognition. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 687–693.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

CSC - IT Center for Science. (1998). The Helsinki Korp Version of the Finnish Text Collection.

Enarvi, S. and Kurimo, M. (2016). Theanolm - an extensible toolkit for neural network language modeling. In *INTERSPEECH*, pages 5; 3052–3056.

Enarvi, S., Smit, P., Virpioja, S., and Kurimo, M. (2017). Automatic speech recognition with very large conversational Finnish and Estonian vocabularies. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2085–2097.

Goodman, J. (2001). Classes for fast maximum entropy training. In *ICASSP*, pages 561–564.

Iskra, D. J., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F., and Kießling, A. (2002). Speecon - speech databases for consumer devices: Database specification and validation. In *LREC*. European Language Resources Association.

Kim, Y., Gao, Y., and Ney, H. (2019). Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1246–1257. Association for Computational Linguistics.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *ICASSP*, volume 1, pages 181–184.

Le, H. S., Oparin, I., Messaoudi, A. K., Allauzen, A., Gauvain, J. L., and Yvon, F. (2011). Large vocabulary SOUL neural network language models. In *INTERSPEECH*, pages 1469–1472.

Mansikkaniemi, A., Smit, P., and Kurimo, M. (2017). Automatic construction of the Finnish parliament speech corpus. In *INTERSPEECH*, pages 3762–3766.

Meister, E., Meister, L., and Metsvahi, R. (2012). New speech corpora at IoC. In *XXVII Fonetiikan päivä — Phonetics Symposium*, pages 30–33.

Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *AISTATS*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *ASRU*. IEEE Signal Processing Society, December.

Rosti, A., Rämö, A., Saarelainen, T., and Yli-Hietanen, J. (1998). Speechdat Finnish database for the fixed telephone network. Technical report, Tampere University of Technology.

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

Smit, P., Gangireddy, S. R., Enarvi, S., Virpioja, S., and Kurimo, M. (2017). Character-based units for unlimited vocabulary continuous speech recognition. In *ASRU*, pages 149–156.

Smit, P., Virpioja, S., and Kurimo, M. (2018). Advances in subword-based hmm-dnn speech recognition across languages. Technical report, Aalto University.

Zhuang, X., Ghoshal, A., Rosti, A., Paulik, M., and Liu, D. (2017). Improving DNN bluetooth narrowband acoustic models by cross-bandwidth and cross-lingual initialization. In *INTERSPEECH*, pages 2148–2152.