Rouhe, Aku; Van Camp, Astrid; Singh, Mittul; Van Hamme, Hugo; Kurimo, Mikko

An Equal Data Setting for Attention-Based Encoder-Decoder and HMM/DNN Models: A Case Study in Finnish ASR

Published in:
Speech and Computer - 23rd International Conference, SPECOM 2021, Proceedings

DOI:
10.1007/978-3-030-87802-3_54

Published: 01/01/2021

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Rouhe, A., Van Camp, A., Singh, M., Van Hamme, H., & Kurimo, M. (2021). An Equal Data Setting for Attention-Based Encoder-Decoder and HMM/DNN Models: A Case Study in Finnish ASR. In A. Karpov, & R. Potapova (Eds.), *Speech and Computer - 23rd International Conference, SPECOM 2021, Proceedings* (pp. 602-613). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 12997 LNAI). Springer. https://doi.org/10.1007/978-3-030-87802-3_54

An equal data setting for Attention-based Encoder-Decoder and HMM/DNN models: A case study in Finnish ASR

Aku Rouhe[1], Astrid Van Camp[2], Mittul Singh[1], Hugo Van Hamme[2], Mikko Kurimo[1]

[1] Aalto University, Department of Signal Processing and Acoustics
{aku.rouhe,mikko.kurimo}@aalto.com
[2] KU Leuven, Department of Electrical Engineering
hugo.vanhamme@esat.kuleuven.be

Abstract. Standard end-to-end training of attention-based ASR models only uses transcribed speech. If they are compared to HMM/DNN systems, which additionally leverage a large corpus of text-only data and expert-crafted lexica, the differences in modeling cannot be disentangled from differences in data. We propose an experimental setup, where only transcribed speech is used to train both model types. To highlight the difference that text-only data can make, we use Finnish, where an expert-crafted lexicon is not needed. With 1500h equal data, we find that both ASR paradigms perform similarly, but adding text data quickly improves the HMM/DNN system. On a smaller 160h subset we find that HMM/DNN models outperform AED models.

Keywords: HMM/DNN, Attention-based Encoder-Decoder, Equal data

A large part of recent speech recognition (ASR) approaches can be divided into two categories: end-to-end attention-based encoder-decoder models (AED) and hybrid hidden Markov model deep neural network (HMM/DNN) models. There are many reasons to choose one approach or the other: some reasons are theoretical (e.g. emphasising the joint optimization of AED models), some practical (e.g. needing the phone-level alignments provided by HMM/DNN models), some simply empirical (which has lower word error rate). The diverged approaches have naturally been compared in terms of performance. Performance comparisons are inherently empirical and the results depend on the constraints of the task [3, 17]. These constraints generally mean the data that is available, as well as any technical limitations such as a maximum latency. Without special techniques [9, 27, 10, 20], end-to-end training only uses transcribed speech, but it is typical for standard ASR tasks to also include expert-crafted pronunciation dictionaries and a large amount of text-only data, which standard HMM/DNN models can leverage.

In this work we focus on the text resource constraints. From a practical perspective, it is natural to include extra text in standard tasks: text-only data

is usually much more plentiful or cheap to produce and thus available to help. From an academic perspective, we argue that experiments with different data constraints could disentangle the effect of differences in data from differences in modeling.

Firstly we propose an experimental design where the HMM/DNN approach is constrained to use the same data as an end-to-end trained AED model. The experimental design allows us to disentangle the effect of extra text data from model performance. In this constrained setting, the focus shifts to building a suitable language model for the HMM/DNN despite the lack of extra text data, rather than trying to augment the AED approach with extra text data.

Secondly, we find that under this design, in a 1500h Finnish ASR task, the AED and HMM/DNN approaches perform similarly, but on a smaller 10% data task, the HMM/DNN model outperforms the AED model. We find Finnish especially well suited to this constraint, because Finnish has a very transparent orthography, and thus the effect having an expert-crafted lexicon (or lack thereof) is annulled.

Thirdly, we extend the subword lexicon handling to support SentencePiece models for HMM/DNNs, as we find subword models vital for the constrained HMM/DNN.

1 Related work

Using an external language model with the AED approach in shallow fusion [9] can also achieve an equal data setting, but the resulting system is no longer trained end-to-end. Using more audio data is shown to reduce dependency on an external language model [28].

There are many special methods which extend end-to-end training of AED models so that text-only data can be used, including tighter integration of a language model [27], and synthesizing audio-side information for text-only data [10, 20].

Comparisons between AED and HMM/DNN approaches exist, but to the best of our knowledge the equal data experimental design proposed here has not been explored before. It is generally thought that with large data sets, AED performs as well or even better than HMM/DNN [3], and with less data HMM/DNN starts to fare better [30], and that both approaches can be certainly be competitive [17, 8].

2 Data

Our training dataset is derived from a combination of three speech datasets: the large Finnish parliament dataset (≈1560h) containing recordings from the Finnish parliments sessions [18], the Speecon corpus (≈160h) containing read speech in various conditions [12], and the Speechdat database (≈220h) containing read and spontaneous, phonetically rich telephone quality (8kHz) speech from

a large number of speakers [21]. After the Kaldi toolkit [19] standard cleanup, a combined ≈1500h training set remains. From this data, 10% of utterances (randomly sampled) are taken to form a ≈160h smaller scale training set.

The transcripts of the full training set consists of ≈9M words, with ≈400k unique words. The 10% subset transcrips have ≈900k words, with ≈100k unique words. In the equal data setting, only the transcripts are used for language modeling. However, there is an order-of-magnitude larger text-only dataset, the Kielipankki text corpus [5] available, consisting of newspaper articles and books. The Kielipankki corpus has a total of ≈143M words, with ≈4.2M unique words.

The main evaluation data is broadcast news shows from Finland's national public service media company YLE. The test set is ≈6h. There is a separate YLE development set of the same type, of ≈5h. However, no matching training set exists for this data. It can be expected that particularly the news articles in the Kielipankki corpus can be helpful on the YLE data. The Finnish parliament data also has a test set, which is in-domain for the training data. It has two sections: one for speakers seen in the training data, one for unseen speakers. Surprisingly, empirically the unseen speakers test data has been *easier*; probably the members of parliament that speak rarely more often read their speech from notes.

3 Attention-based system

We train AED models end-to-end using the ESPnet toolkit [29]. In initial experiments, Transformer architectures outperformed RNN-based architectures, thus we opt for a Transformer model. This architecture also generally fares well in AED applications [13]. The ESPnet toolkit supports using the CTC criterion in a multi-task setup (with weight α) to aid the encoder training [14]. The separate CTC decoder is can also be used in decoding (with weight γ). We find that it offers a minor improvement in this task.

The common output units for AED models are characters or larger subword units. We conduct experiments with SentencePiece [15] BPE subword units, as they are well integrated into ESPnet, but interestingly find no improvement in this task. Thus we simply use character output units.

For the full 1500h data we tune the model size on the YLE development set, ending with 16 encoder layers and 8 decoder layers of width 2048 each. We also search for ideal weights for the CTC multitask learning and decoding, setting $\alpha = 0.2$, $\gamma = 0.1$. Besides output units, CTC parameters, and model size, we refer to recipes of similarly sized data for the other hyperparameters, using dropout rate 0.1 and label smoothing 0.1. We train the models with early stopping on the YLE development data. Before decoding, we average the last 10 model checkpoints' weights, and decode with the averaged model.

On the smaller 10% data we simply used the same hyperparameters as with the 1500h model. We experimented with a smaller model size (anticipating that having less data could lead to overfitting), but the same model size performed better.

4 HMM system

For the HMM/DNN approach, we use the Kaldi toolkit. As a baseline for acoustic model development we refer to results from [26]. Like [26], we use context- and word-position dependent grapheme-units. However, their models use i-vectors for speaker-aware training, as is typical in the Kaldi toolkit. The AED approach seems to get similar benefit from speaker-aware training as the HMM/DNN methods [6, 22, 1], and so we argue that either both or neither of the paradigms should use speaker-aware methods. For simplicity, we opt for neither.

We train a new, large (chain-style) time delay neural network (TDNN) based model, without i-vectors. By using more TDNN-layers than Smit et al. we are able to achieve similar performance on the YLE development set without i-vectors, as shown in table 1. Thus we choose this acoustic model for the HMM approach. On the smaller 10% data, we simply keep the same hyperparameters.

Table 1. Validating the new acoustic model, without i-vectors, against the best published result with TDNN-BLSTM and i-vectors.

Model	YLE Dev WER
TDNN-BLSTM + i-vectors	17.3
Large TDNN	17.4

The Finnish language is agglutinative, which leads to very large vocabularies. Particularly on small text datasets, such as the speech transcripts used here, the data sparsity problem is evident, if using traditional word-based language models. Instead, we use a subword language model, which dramatically reduces the vocabulary size. With subword language models, long n-gram contexts may be necessary for good performance [11], and therefore we use the variKN toolkit [23] to train Kneser-Ney varigram n-gram models of up to 10-gram order. Additionally, we train standard LSTM-based RNNLMs with TheanoLM [7], optimizing network size on the YLE development data transcript perplexity. This leads to network size 1024 units. The RNNLM is applied in lattice rescoring.

To segment the training transcripts into subwords, we use the same SentencePiece BPE algorithm as in the AED approach. This way, both approaches had the same segmentation style available. We optimize the variKN scale parameter, and the number of BPE units, for perplexity on the YLE development set transcripts. Table 2 shows the resulting numbers of BPE merges for different model types and data sizes. We optimize the language model weight for the YLE development set word error rate.

4.1 Correct subword handling in Lexicon FST

Using subword-based language models requires some care in constructing the lexicon FST. Firstly, when using word-position-dependent units [16], the lexicon should take subword concatenation into account when handling word-boundary

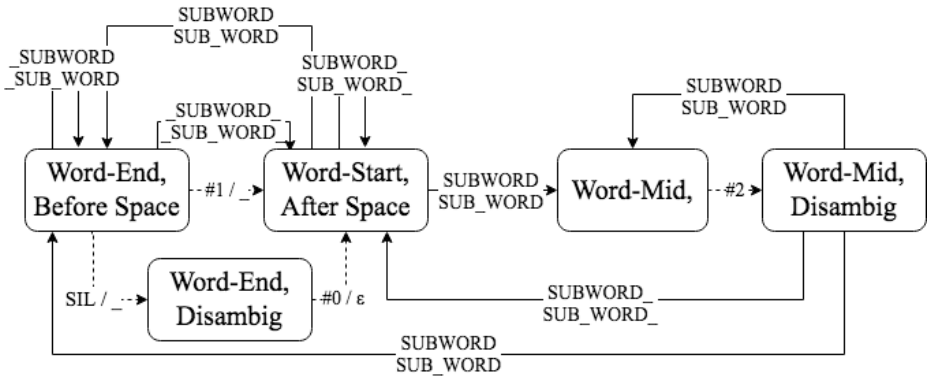

Fig. 1. SentencePiece lexicon FST showing which word positions SentencePiece units can be placed at, depending on where the space character ▁ appears on example unit `SUBWORD` (for instance, at the start of the unit: `▁SUBWORD`). Also shows three disambiguation symbols, and two more are needed for spaces inside units.

units [25]. For example, the lexicon should only transduce the word-start and word-end units at true word boundaries, but not at every subword boundary. Secondly, as SentencePiece units may have word boundaries as part of the subword units, the optional silence at word boundaries needs to be handled specially. As part of this work, we extend the tools introduced by [25] to support Sentencepiece segmentation[3].

Figure 1 shows which different word positions SentencePiece units can be placed at, depending on where the space character appears in them. The figure also shows three word-position dependant disambiguation symbols; two more are needed for in-unit spaces.

To validate the proposed SentencePiece lexicon FST, we compare the ASR results to those of Morfessor generated subwords [4, 24] as presented by [25], as shown in table 3. The SentencePiece models achieve a slightly lower error rate on the larger corpus and slightly higher on the smaller one. The SentencePiece lexicon comes at the cost of larger search graph (HCLG) FSTs compared to Morfessor, for a similar sized language model, because SentencePiece requires a more complex word boundary handling scheme.

5 Results

Table 4 lists the model sizes. Everything combined, the 1500h HMM/DNN system matches the AED model size quite well.

Table 5 shows the performance of the best equal data AED and HMM/DNN systems on the YLE evaluation data. The absolute word error rates are quite similar on the full 1500h data. Note that the percentage in brackets indicates

[3] `https://github.com/aalto-speech/subword-kaldi`

Table 2. Optimal number of BPE merges for different data sizes and model types. In this task the AED systems did not benefit from BPE units, and character-level models were used instead. This was optimized on YLE Development WER (end-to-end). On the HMM side, the number of units was optimized for LM perplexity. With the limited 10% transcripts, a small number of BPE units was marginally better than character units, but with the full transcript and with the ≈17 times larger Kielipankki text data, BPE units clearly improved.

Model	Number of merges
HMM 10% Transcripts	≈100
HMM All Transcripts	≈1700
HMM Kielipankki	≈10000
AED 10% Data	0
AED All Data	0

Table 3. SentencePiece units compared to Morfessor units on the YLE development data, using the same acoustic models. The *small* models used the Parliament transcripts, while the *large* models used the larger Kielipankki corpus.

Model	YLE Dev WER
Morfessor_{small}	27.4
$\text{SentencePiece}_{small}$	28.4
Morfessor_{large}	17.4
$\text{SentencePiece}_{large}$	15.8

the bootstrap estimate [2] of the probability of improving over the competing model (with 10000 repetitions); with strict 95% confidence cutoffs the YLE 1500h WER result would be inconclusive. On the 10% data, the HMM/DNN system outperforms the AED system.

Since Finnish tends to have long words with inflections, character error rate (CER) is a useful metric as well. Table 6 shows the YLE CER evaluation. On the 1500h data this paints a different picture than the WER evaluation: the AED model outperforms the HMM/DNN system. However, on the 10% data the HMM/DNN system has lower CER as well. This CER evaluation did not consider spaces.

The equal data results of tables 5 and 6 are contrasted by the results in table 7, which details a set of experiments where a number of random lines was added from the Kielipankki corpus to the HMM/DNN language model training data set (in addition to the speech transcripts). This shows how fast the WER would decrease as slightly better matching text data was added. New language models were trained on the total resulting text data. We optimized the language model and subword segmentation parameters again for each new text corpus obtained this way.

The equal data models from tables 5 and 6 are also evaluated on the Finnish Parliament test sets. Table 8 shows the WER results and table 9 shows the CER results. We emphasize that these models were optimized on the YLE Development data. For constrast, we show a result [18] that was optimized on the

Table 4. Comparison of model capacity in terms of number of parameters. N-gram number of parameters measured by number N-grams included in the model; the number of arcs in the resulting finite state transducer is about 50% more.

Model	Number of parameters
AED	35M
HMM Total	35M
Acoustic model	20M
N-gram LM	6M
RNNLM	9M
$AED_{10\%}$	35M
$HMM_{10\%}$ Total	22M
Acoustic model	20M
N-gram LM	2M

Table 5. At the top, the comparison of the best attention (AED) and HMM based models, trained on the 1500h equal data, evaluated on the YLE broadcast data. The percentage in brackets indicates the bootstrap estimate [2] of the probability of improving over the competing model (inside the same horizontal lines). In the middle: the RNNLM (also trained on transcripts only) does not meaningfully improve over the 10-gram Kneser-Ney LM. In the bottom, similar comparison of the best attention and HMM-based models trained on the smaller 10% random subset of the 1500h equal data.

YLE		
Model	Dev WER	Test WER
AED	28.7	**27.8** (78%)
HMM	**28.4** (87%)	28.1
HMM +RNNLM	29.0	28.0
$AED_{10\%}$	36.8	35.8
$HMM_{10\%}$	**35.0** (100%)	**34.0** (100%)

Table 6. Character error rate equivalent of table 5

YLE		
Model	Dev CER	Test CER
AED	**5.96** (95%)	**5.57** (99.7%)
HMM	6.13	5.88
HMM +RNNLM	6.29	6.07
$AED_{10\%}$	7.73	7.21
$HMM_{10\%}$	**7.24** (100%)	**6.96** (100%)

Parliament data: this is by far the lowest WER. Based on both WER and CER, the 1500h equal data AED and HMM/DNN models perform similarly. Again, on the 10% data, the HMM/DNN system outperforms the AED system.

Table 7. YLE evaluation WER decreasing as more text data is added to the HMM/DNN approach from the Kielipankki corpus. The number in brackets indicates how large the resulting text dataset is compared to the original.

YLE

Text data condition	Dev WER	Test WER
Transcripts only	28.4	28.1
+100k KP lines (1.1x)	23.7	23.7
+200k KP lines (1.2x)	22.4	22.1
+500k KP lines (1.7x)	20.3	20.5
+1M KP lines (2.3x)	18.8	19.1
All of KP (16.9x)	15.8	16.2

Table 8. At the top, as a baseline we show Parl-opt HMM, which uses a large (20M token) in-domain text corpus for language modeling, and is optimized for the parliament data. Then, the same models as table 5, evaluated on the Parliament test set (but models still optimized for YLE). Here the RNNLM improves slightly, indicating it has slightly overfit to the Parliament data compared to YLE data in table 5.

FINNISH PARLIAMENT

Model	Test-Seen WER	Test-Unseen WER
Parl-opt HMM[18]	5.9	5.2
AED	**10.2** (95%)	10.2
HMM	10.6	**9.6** (98%)
HMM +RNNLM	**9.7**	**9.0**
$\text{AED}_{10\%}$	18.2	18.1
$\text{HMM}_{10\%}$	**14.9** (100%)	**13.7** (100%)

Table 9. Character error rate equivalent of table 8. Character error rates were not available for the Parl-opt HMM.

FINNISH PARLIAMENT

Model	Test-Seen CER	Test-Unseen CER
AED	**2.91** (81%)	**2.99** (56%)
HMM	2.99	3.01
HMM +RNNLM	**2.80**	**2.85**
$\text{AED}_{10\%}$	4.64	4.92
$\text{HMM}_{10\%}$	**3.66** (100%)	**3.69** (100%)

6 Discussion

On the whole it can be said that in the full 1500h equal data scenario, the AED and HMM/DNN approaches performed similarly. The CER results on the YLE data favour the AED approach, which may be connected to the observation that character-level units performed better than BPE-subwords for the AED approach, and generally CER evaluation de-emphasizes the role of the language model.

In contrast to the full 1500h data, on the 10% utterances (160h) subset, the HMM/DNN consistently performed better than the AED approach. This is in line with the general understanding that HMM/DNN systems fare better with less data.

Table 7 shows just how much the extra text-only data could help. Text-only data can be naturally leveraged by HMM/DNN models, but improvements could certainly be found with the AED text-data leveraging techniques as well. Which of these techniques works best for AED models is an active research question, but not in the scope of this work. Together these experiments show how the proposed equal data experimental design can provide a new perspective: with the data being equal, the models' inherent strengths and weaknesses are emphasized.

We used our best transformer model to represent the AED approach and the HMM/DNN approach was represented by a strong TDNN acoustic model coupled with our best optimized language model. The models have similar number of parameters. Undoubtedly both approaches could have been optimized further: on competitive datasets such as LibriSpeech, the research community competes for tenths of percentage points of WER. Thus we feel that the differences between AED and HMM/DNN models on the full 1500h data should be seen as inconclusive: both performed similarly. However, some results were clearer: the 10% data favors the HMM/DNN approach and the additional text-only data was obviously the key to better error rates on the YLE data.

Thus the results here suggest there is a medium dataset scale, where the AED and HMM/DNN can perform similarly, as long as techniques for leveraging external text-only data are developed for the AED-models.

7 Conclusions

We have proposed an *equal data* experimental design where HMM/DNN systems and end-to-end trained attention-based systems are compared by artificially limiting the data to transcribed speech only. This sort of task is well suited for Finnish, which has a transparent orthography, ruling out the lexicon's effect. In our equal data experiments, HMM/DNN models and AED models perform similarly at full 1500h data, but HMM/DNN models outperform AED models on a smaller 10% subset. Extra text-only data clearly improves the HMM/DNN models, which emphasizes our argument that when HMM/DNN models and end-to-end trained AED models are compared, unequal data resources can hide differences in modeling.

As part of developing the HMM/DNN models under constraints set by this comparison, we also implement a lexicon FST which handles the SentencePiece subword segmentation correctly.

Acknowledgements This work was supported by EU's Horizon 2020 research and innovation programme via the project MeMAD (GA 780069). The computational resources were provided by Aalto ScienceIT.

References

1. Bansal, S., Malhotra, K., Ganapathy, S.: Speaker and language aware training for end-to-end asr. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 494–501 (2019). https://doi.org/10.1109/ASRU46091.2019.9004000
2. Bisani, M., Ney, H.: Bootstrap estimates for confidence intervals in asr performance evaluation. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 1, pp. I–409 (2004). https://doi.org/10.1109/ICASSP.2004.1326009
3. Chiu, C., Sainath, T.N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R.J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., Bacchiani, M.: State-of-the-art speech recognition with sequence-to-sequence models. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4774–4778 (2018)
4. Creutz, M., Lagus, K.: Unsupervised discovery of morphemes. In: Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning. pp. 21–30. Association for Computational Linguistics (Jul 2002). https://doi.org/10.3115/1118647.1118650, https://www.aclweb.org/anthology/W02-0603
5. CSC - IT Center for Science: The helsinki korp version of the Finnish text collection, url: http://urn.fi/urn:nbn:fi:lb-2016050207 (1998), http://urn.fi/urn:nbn:fi:lb-2016050207
6. Delcroix, M., Watanabe, S., Ogawa, A., Karita, S., Nakatani, T.: Auxiliary feature based adaptation of end-to-end asr systems. In: Proc. Interspeech 2018. pp. 2444–2448 (2018). https://doi.org/10.21437/Interspeech.2018-1438, http://dx.doi.org/10.21437/Interspeech.2018-1438
7. Enarvi, S., Kurimo, M.: Theanolm — an extensible toolkit for neural network language modeling. In: Interspeech 2016. pp. 3052–3056 (2016). https://doi.org/10.21437/Interspeech.2016-618, http://dx.doi.org/10.21437/Interspeech.2016-618
8. Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R.: Conformer: Convolution-augmented Transformer for Speech Recognition. In: Proc. Interspeech 2020. pp. 5036–5040 (2020). https://doi.org/10.21437/Interspeech.2020-3015, http://dx.doi.org/10.21437/Interspeech.2020-3015
9. Gulcehre, C., Firat, O., Xu, K., Cho, K., Bengio, Y.: On integrating a language model into neural machine translation. Computer Speech & Language **45**, 137 – 148 (2017). https://doi.org/https://doi.org/10.1016/j.csl.2017.01.014, http://www.sciencedirect.com/science/article/pii/S0885230816301395
10. Hayashi, T., Watanabe, S., Zhang, Y., Toda, T., Hori, T., Astudillo, R., Takeda, K.: Back-translation-style data augmentation for end-to-end asr. In: 2018 IEEE Spoken Language Technology Workshop (SLT). pp. 426–433 (2018). https://doi.org/10.1109/SLT.2018.8639619
11. Hirsimaki, T., Pylkkonen, J., Kurimo, M.: Importance of high-order n-gram models in morph-based speech recognition. IEEE Transactions on Audio, Speech, and Language Processing **17**(4), 724–732 (2009)
12. Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F., Kiessling, A.: SPEECON – speech databases for consumer devices: Database specification

and validation. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02). European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain (May 2002), http://www.lrec-conf.org/proceedings/lrec2002/pdf/177.pdf
13. Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N.E.Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T., Zhang, W.: A comparative study on transformer vs rnn in speech applications. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 449–456 (2019). https://doi.org/10.1109/ASRU46091.2019.9003750
14. Kim, S., Hori, T., Watanabe, S.: Joint ctc-attention based end-to-end speech recognition using multi-task learning. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4835–4839 (March 2017). https://doi.org/10.1109/ICASSP.2017.7953075
15. Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018)
16. Lee, C.H.: Improved acoustic modeling for speaker independent large vocabulary continuous speech recognition. In: [Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing. pp. 161–164 vol.1 (1991). https://doi.org/10.1109/ICASSP.1991.150302
17. Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer, A., Schlüter, R., Ney, H.: RWTH ASR Systems for LibriSpeech: Hybrid vs Attention. In: Proc. Interspeech 2019. pp. 231–235 (2019). https://doi.org/10.21437/Interspeech.2019-1780, http://dx.doi.org/10.21437/Interspeech.2019-1780
18. Mansikkaniemi, A., Smit, P., Kurimo, M.: Automatic construction of the finnish parliament speech corpus. In: Proc. Interspeech 2017. pp. 3762–3766 (2017). https://doi.org/10.21437/Interspeech.2017-1115, http://dx.doi.org/10.21437/Interspeech.2017-1115
19. 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.: The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (Dec 2011), iEEE Catalog No.: CFP11SRW-USB
20. Rossenbach, N., Zeyer, A., Schlüter, R., Ney, H.: Generating synthetic audio data for attention-based speech recognition systems. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7069–7073 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053008
21. Rosti, A., Rämö, A., Saarelainen, T., Yli-Hietanen, J.: Speechdat finnish database for the fixed telephone network. Tampere University of Technology, Tech. Rep (1998)
22. Rouhe, A., Kaseva, T., Kurimo, M.: Speaker-aware training of attention-based end-to-end speech recognition using neural speaker embeddings. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7064–7068 (2020)
23. Siivola, V., Hirsimaki, T., Virpioja, S.: On growing and pruning kneser–ney smoothed n-gram models. IEEE Transactions on Audio, Speech, and Language Processing **15**(5), 1617–1624 (2007). https://doi.org/10.1109/TASL.2007.896666
24. Smit, P., Virpioja, S., Grönroos, S.A., Kurimo, M.: Morfessor 2.0: Toolkit for statistical morphological segmentation. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 21–24 (2014)

25. Smit, P., Virpioja, S., Kurimo, M.: Improved subword modeling for wfst-based speech recognition. In: Proc. Interspeech 2017. pp. 2551–2555 (2017). https://doi.org/10.21437/Interspeech.2017-103, http://dx.doi.org/10.21437/Interspeech.2017-103
26. Smit, P., Virpioja, S., Kurimo, M.: Advances in subword-based hmm-dnn speech recognition across languages. Computer Speech & Language **66**, 101158 (2021). https://doi.org/https://doi.org/10.1016/j.csl.2020.101158, http://www.sciencedirect.com/science/article/pii/S0885230820300917
27. Sriram, A., Jun, H., Satheesh, S., Coates, A.: Cold fusion: Training seq2seq models together with language models. In: Proc. Interspeech 2018. pp. 387–391 (2018). https://doi.org/10.21437/Interspeech.2018-1392, http://dx.doi.org/10.21437/Interspeech.2018-1392
28. Synnaeve, G., Xu, Q., Kahn, J., Likhomanenko, T., Grave, E., Pratap, V., Sriram, A., Liptchinsky, V., Collobert, R.: End-to-end asr:from supervised to semi-supervised learning with modern architectures. In: ICML 2020 Workshop on Self-supervision in Audio and Speech (2020)
29. Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., Ochiai, T.: ESPnet: End-to-end speech processing toolkit. In: Proceedings of Interspeech. pp. 2207–2211 (2018). https://doi.org/10.21437/Interspeech.2018-1456, http://dx.doi.org/10.21437/Interspeech.2018-1456
30. Zhou, W., Michel, W., Irie, K., Kitza, M., Schlüter, R., Ney, H.: The rwth asr system for ted-lium release 2: Improving hybrid hmm with specaugment. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7839–7843 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053573