Tandem Long-Short Duration-based Modeling for Automatic Speech Recognition

Dalai Mengke,Yan Meng and Péter Mihajlik
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
kedalai.meng@edu.bme.hu, yan.meng@edu.bme.hu, mihajlik.peter@vik.bme.hu

Abstract

This study outlines our duration-dependent modeling experiments on limited-resource Hungarian speech recognition tasks. As it is well known, very short utterances pose significant challenges in automatic speech recognition due to the lack of context and other phenomena. In particular, we found that that the exclusion of shorter speech samples from fine-tuning for longer duration test data significantly improves the recognition rate measured on public Hungarian datasets, BEA-Base and CommonVoice (CV). Therefore we apply a tandem modeling approach, separate models are used for short and long duration test data. Our strategy improved the ability to recognize short utterances while maintaining recognition of long utterances efficiently, which led to a significant increase in overall recognition accuracy.

Keywords: automatic speech recognition, short utterance, duration dependent modeling, transfer learning

1. Introduction

End-to-end deep neural approach (Graves and Jaitly, 2014) and transfer learning have been proven to be effective techniques (Kunze et al., 2017) (Huang et al., 2020) used widely for automatic speech recognition (ASR). Transfer learning allows for a swift transition from a pre-trained model to another speech recognition model, often more effective than training from scratch and can be considered as best practice in low-resource tasks. This study, however, has identified a significant phenomena when testing ASR models trained in such a way. It is shown on Figure 1 that utterances in the test set with higher error rates tend to be shorter. As can be seen from Table 1, after removing a small number of shorter test samples from the test set, the recognition accuracy of the remaining test set became significantly higher. Obviously, the standard ASR approach still has limitations in processing short utterances. Further research might be needed to improve recognition accuracy for short utterances, thereby enhancing the comprehensive performance of speech recognition systems.

The phenomenon of degraded accuracy for shorter chunks may be due to a combination of factors. First, short utterances in speech recognition often contain less substantive information and naturally, the context is reduced, which poses a challenge both for training and for accurate recognition. Second, there is a potential bias in model training: if long utterances dominate the training data, the model may perform poorly in recognizing short utterances.

Based on these observations, although models fine-tuned based on transfer learning perform well in recognizing long utterances, there is room for improvement with respect to process short utterances. This study proposes a hypothesis: developing a model specifically for short utterances and using it in parallel with the existing model after transfer learning for long utterances might improve the overall recognition effect. This approach would combine the advantages of both models, i.e., the efficient recognition ability of long utterances and the specialized processing capability optimized for short utterances, aiming to achieve more comprehensive and accurate speech recognition performance. Future research could explore the effectiveness of this dual-model parallel strategy and how to optimize models to provide the best recognition performance for utterances of different duration.

Recent research advancements reveal that adaptation technology can be an effective alternative to traditional transfer learning, with significant advantages in speed and efficiency (Houlsby et al., 2019), while achieving comparable performance (Thomas et al., 2022). Based on this finding, this paper proposes a tandem model methodology, which is to further fine-tune short utterances using adaptation technology on model which has been already fine-tuned through transfer learning. This approach aims the model to improve its recognition capability for short utterances while maintaining good performance for long utterances.

Overall, this study will explore the effectiveness of adaptation technology in enhancing the recognition performance of short utterances in automatic speech recognition. Through this method, we expect to propose a more precise and efficient au-

SIGUL2024 Workshop, pages 331–336
21-22 May, 2024.

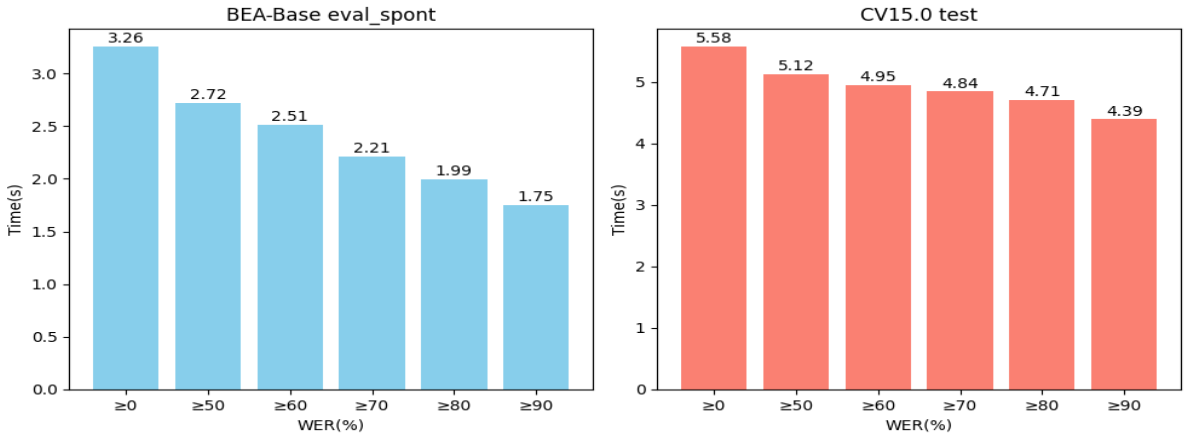

Figure 1: These two bar charts show the error rate vs utterance duration in a tested transcript of two Hungarian-language models obtained by transfer learning from an English pre-trained model. The left and right bar charts show the relationship between error rate and utterance duration in the test sets of BEA-Base and CV, respectively, where the vertical coordinate is time and the horizontal coordinate is the Word Error Rate (WER). The WER $\geq 0\%$ refers to that the average utterance duration of the entire test set. And ≥ 50 refers to the average duration of all the utterance with a WER $\geq 50\%$ in the test transcripts, etc.

tomatic speech recognition system, especially in handling language inputs of varying duration.

Duration	BEA-Base (eval-spont)(%)
$T \geq 0s$	25.42
$T \geq 2.0s$	24.85
$T \geq 2.5s$	24.70
$T \geq 3.0s$	24.72
$T \geq 3.5s$	24.65

Table 1: This table shows the change in the word error rate (WER) of the test after excluding some of the shorter utterances from BEA-Base's test set (enval-spont). $T \geq 0s$ means that no data from the test set is excluded, i.e., the entire test set is used, $T \geq 2.0s$ means that the test is performed with samples of 2 seconds and more, etc.

2. Relationship Between Utterance Duration and Error Rate

Here we explore the relationship between utterance duration and error rate, and we use two different datasets, BEA-Base (Mihajlik et al., 2022a) and CommonVoice (CV) (Ardila et al., 2019), and conduct experiments in the Conformer (Gulati et al., 2020) modeling framework. Both models were transferred from an English pre-trained model (STT En Conformer-CTC Small[1]) to Hungarian and were trained with their respective training sets and tested with their test sets.

[1] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_small

During the test phase, we evaluated the test set on each dataset to observe the performance of the models trained by the respective training sets. After the test was completed, we filtered out the samples with word error rates (WER) higher than 0.5, and for these samples, we plotted histograms of sample duration versus error rate for different error rate thresholds. The results show the higher the error rate threshold, the shorter the average duration of the utterances in both the BEA-Base (Mihajlik et al., 2022a) and CommonVoice (Ardila et al., 2019) datasets.

However, a significant improvement in the accuracy of the test was then found when doing the test on the first fine-tuned model with samples below a certain duration threshold(s) removed. Here the treatment was done on two separate datasets, Table 1 shows the results of applying this operation on BEA-Base (eval-spont), and Table 2 shows the results of applying the same operation on CV15.0 test.

3. Methodology

3.1. Initial Fine-tuning

The first step of the method is to fine-tune an English pre-trained model (STT En Conformer-CTC Small) to the speech recognition task in Hungarian using a transfer learning approach. This process involves applying the pre-trained model to a corpus of the target language (Hungarian) and optimizing the model parameters through fine-tuning with a view to obtaining a model that recognizes Hungarian.

Duration	CV15.0 (test)(%)
$T \geq 0s$	23.72
$T \geq 3.0s$	23.62
$T \geq 3.5s$	23.47
$T \geq 4.0s$	23.33
$T \geq 4.5s$	23.25
$T \geq 5.0s$	23.02
$T \geq 5.5s$	23.05
$T \geq 6.0s$	23.00
$T \geq 6.5s$	22.96

Table 2: This table shows the change in the word error rate (WER) of the test after excluding some of the shorter utterances from CommonVoice's test set (CV15.0 test). The $T \geq 0s$ refers to no data from the test set is excluded, i.e., the entire test set is used, $T \geq 3.0s$ refers to the test is performed with utterances duration ≥ 3 seconds, etc.

3.2. Model Fine-Tuning by Short Utterances

During the training and validation phases, a threshold T is set based on the duration of the speech samples, dividing them into long and short utterances. For short utterances, adaptation technique is used to further fine-tune the transferred model. Adapter layers are embedded into the initial fine-tuned model, specifically training by short utterance samples to enhance the model's performance in recognizing short utterances.

4. Experimental Set-up

4.1. Common Setting

In this study, the hardware configuration consists of a system equipped with dual Nvidia A6000 graphics cards, ensuring efficient processing capabilities for deep neural network training and inference. The model chosen for this investigation is the Conformer Small model (Gulati et al., 2020), renowned for its effectiveness in speech recognition tasks. The linear adapter (NVIDIA, 2024) was applied for fine-tuning the model with short utterances.

Regarding hyper-parameter settings, a learning rate of 0.002 is applied to optimize the training process, a batch size of 32 is used, coupled with the utilization of Connectionist Temporal Classification (CTC) loss function (Graves et al., 2006). This loss function is particularly suited for sequence-to-sequence problems typical in speech recognition.

To facilitate the experiments, the NVIDIA NeMo toolkit (Kuchaiev et al., 2019), version 1.22.0, is employed. This toolkit is widely recognized for its robust features in speech and language processing. For all other parameters not mentioned, NeMo's default recipe is used.

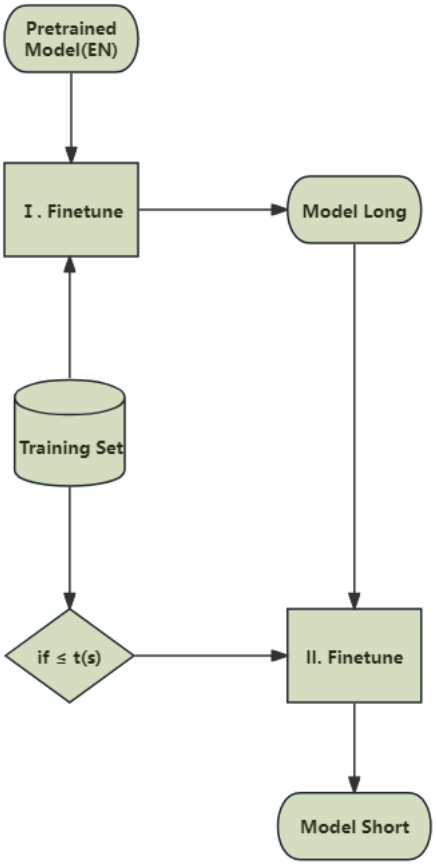

Figure 2: Workflow. The figure shows the training process for both long and short models. Firstly, a pre-trained model in English and the entire dataset was fine-tuned to obtain the model for long utterances(M_L), and then utterances of duration < T, i.e., short utterances, short utterances, were identified from this dataset and fine-tuned again using short utterances to obtain the model for short utterances (M_S).

4.2. Transfer Learning Phase

Here, an English pre-trained model is used initially and then it is fine-tuned on two Hungarian datasets (BEA-Base and CV15.0). The fine-tuning process consists of training the models on the BEA-Base and CV datasets for 200 and 100 epochs, respectively, which results in two different models specifically adapted to each dataset. After the training phase, these models will be evaluated on their respective test datasets. The purpose of the evaluation is to establish a baseline error rate, which serves as a benchmark for the performance of dual-model.

Split by Time(s)	ER_S on M_L(%)	ER_S on M_S(%)
2.5	26.30	25.51
3.0	25.67	25.10
3.5	25.63	25.12

Table 3: This table shows the results of testing eval-spont on M_L and M_S with short utterance datasets that have been segmented with different thresholds(T), ER_S refers to the error rate of the short dataset. Such a comparison is to demonstrate that the model M_S, fine-tuned with shorter sentences, achieves better recognition of short sentences compared to the initial fine-tuning of the obtained model M_L.

Split by Time(s)	ER_S on M_L(%)	ER_S on M_S(%)
4.5	26.04	25.86
5.0	25.46	24.46
5.5	24.71	23.79
6.0	24.34	23.17
6.5	24.14	22.90

Table 4: This table shows the results of testing CV15.0 test set on M_L and M_S with short utterance datasets that have been segmented with different thresholds(T), ER_S refers to the error rate of the short dataset. Such a comparison is to demonstrate that the model M_S, fine-tuned with shorter sentences, achieves better recognition of short sentences compared to the initial fine-tuning of the obtained model M_L.

T(s)	N_{ErrorL}/N_{WordL}	ER_L on M_L(%)	N_{ErrorS}/N_{WordS}	ER_S on M_S(%)	Av. ER(%)
-	-	-	-	-	25.42(Baseline)
2.5	7083 / 28673	24.70	1660 / 6505	25.51	24.85
3.0	6291 / 25445	24.72	2443 / 9733	25.10	24.82
3.5	5484 / 22241	24.65	3249 / 12937	25.12	**24.82**

Table 5: This is the result of testing on the M_L model using the full BEA-Base's test set(eval-spont), compared with the test results using M_L and M_S working together(Test separately according to duration). It shows that the test set (eval-spont) was segmented into long utterances set, and short utterances set from 2.5 to 3 seconds according to different duration thresholds T. The results of long utterances tested on M_L are labeled as ER_L on M_L, while the results of short utterances tested on M_S are denoted as ER_S on M_S. The average word error rate, $Av.ER$ is computed from Equation 1. Additionally the baseline was only measured directly with the first fine-tuned model using the full test set(eval-spont), so it was not calculated using this formula.

T(s)	N_{ErrorL}/N_{WordL}	ER_L on M_L(%)	N_{ErrorS}/N_{WordS}	ER_S on M_S(%)	Av. ER(%)
-	-	-	-	-	23.72(Baseline)
4.5	16162 / 69513	23.25	3550 / 13726	25.86	23.68
5.0	13786 / 59888	23.02	5712 / 23351	24.46	23.42
5.5	11436 / 49612	23.05	8001 / 33627	23.79	23.35
6.0	8895 / 38658	23.00	10330 / 44581	23.17	23.09
6.5	6775 / 29497	22.96	12310 / 53742	22.90	**22.92**

Table 6: This is the result of testing on the M_L model using the full CV15 test set, compared with the test results using M_L and M_S working together(Test separately according to duration). It shows that the test set (CV15.0 test) was segmented into long utterances set, and short utterances set from 4.5 to 6.5 seconds according to different duration thresholds T. The results of long utterances tested on M_L are labeled as ER_L on M_L, while the results of short utterances tested on M_S are denoted as ER_S on M_S. The average word error rate, $Av.ER$ is computed from Equation 1. Additionally the baseline was only measured directly with the first fine-tuned model using the full test set(CV15.0 test), so it was not calculated using this formula.

4.3. Dataset Segmentation

In this step, a specific time threshold T was set to distinguish between long and short utterances in the dataset. Specifically, utterances with a duration $\geq T$ were classified into a set of long utterances, while those with a duration $< T$ were classified into a set of short utterances. This re-

search involved two different Hungarian language datasets, namely BEA-Base (Mihajlik et al., 2022b) and CV15.0 (Ardila et al., 2019). For the BEA-Base dataset, the threshold T was set between 2.5 to 3.5 seconds for the training set (Train-114), validation set (dev-spont), and test set (eval-spont). For the CV15 dataset, referred to as CV15.0, the T value ranged from 4.5 to 6.5 seconds, applied to the training set (train), validation set (dev), and test set (test). Furthermore, to avoid issues related to limited data amount of short utterances during further fine-tuning, the threshold T for the BEA-Base dataset was set starting from 2.5 seconds, unlike the starting point of 2 seconds as Table 1, the CV15.0 dataset was set starting from 4.5 seconds, unlike the starting point of 3 seconds as Table 2.

4.4. Training Short Utterance Model

We employ the method of embedding adapters into the post-transfer learning model for fine-tuning, which serves to efficiently retain the original model information while also achieving rapid adjustments. Specifically, in the BEA-Base and CV15.0, utterances from the training and validation sets that are shorter than the defined time threshold T, are used for this purpose. The adapter is trained for a duration of 50 epochs, a "linear" type adapter was applied (NVIDIA, 2024).

4.5. Test and Evaluation

After completing the steps described, we have developed two models: M_L, a model fine-tuned for processing longer utterances, and M_S, a model adept at handling short utterances, created by embedding an adapter and performing additional fine-tuning. In the test phase, these two models are employed in a collaborative manner.

For utterances in the test set that are longer than the threshold T, M_L is used to calculate the error rate for long utterances (ER_L). Conversely, for utterances shorter than T, the M_S is utilized to determine the error rate for short utterances (ER_S). This dual-model strategy is designed to optimize speech recognition accuracy across varying utterance duration.

4.6. Combined Accuracy Calculation

In assessing the composite accuracy of a speech recognition model, it is important to consider both the error rates of long utterances (ER_L) and short utterances (ER_S). This evaluation also involves accounting for the number of erroneous words in long utterances, denoted as N_{ErrorL}, and the total number of words in long utterances, represented as N_{WordL}. Similarly, for short utterances, the number of erroneous words, N_{ErrorS}, and the total number of words, N_{WordS}, are also factored into the calculation. The average error rate (Av.ER) is given by Formula 1.

$$Av.ER = \frac{N_{\text{ErrorL}} + N_{\text{ErrorS}}}{N_{\text{WordL}} + N_{\text{WordS}}} \quad (1)$$

5. Results Analysis

The experimental results of this study reveal some key findings. First, as demonstrated in Table 3 and Table 4, for the task of processing short utterances, the model obtained by using the adapter technique to fine-tune it again exhibits a significant performance improvement compared to the model that has only been fine-tuned by initial transfer learning both on BEA-Base and CV15.0. This result shows that the model fine-tuned again by using short utterances has a stronger short utterance recognition ability.

Furthermore, to address the lack of performance of the model fine-tuned with transfer learning using the full dataset for short utterance recognition, this study proposes a two-model strategy that works in tandem. This strategy combines two models: a model that has been fine-tuned by full-parameter transfer learning optimized specifically for long utterances, and a model that has been fine-tuned again for short speech using an adapter technique. With this combination, the two models work together on speech samples of different duration.

The results, shown in Table 5 for BEA-Base, and the results, shown in Table 6 for CV15.0, indicate that when the two models co-work, there is a 2.4% relative boost in WER for BEA-Base, and a 3.2% relative boost on CV15.0 compared to baseline that uses transfer learning and all training set to fine-tune the model.

6. Conclusion

In this paper, it was found that the automatic recognition of short utterances are generally more difficult than long ones. For this challenge, we proposed a tandem modeling approach: separate models are obtained by various fine-tuning steps for short and long utterances and these models work together achieving a noticeable improvement on WER on two publicly available Hungarian datasets (BEA-Base, CV15.0).

However, this tandem model approach has limitations. The added step of determining the length of utterances might lead to delays and other problems in practical applications. Moreover, training the model can be challenging for datasets where the distinction between short and long sentences is not clearly defined.

As for future work, we want to generalize the use of the two-model cooperation strategy across a wider range of datasets as well as a wider range of languages to explore the potential of this approach.

7. Acknowledgment

This research benefited greatly from the support provided by the Hungarian Linguistic Research Center in the development of the BEA-Base dataset. This work was supported partially by NKFIH-828-2/2021 (MILab), by the NVIDIA Academic Hardware Grant and by the NKFIH K143075 and K135038 projects of the NRDI Fund. Thanks are also extended to the Budapest University of Technology and Economics and NVIDIA Academic Hardware Grant for their vital contribution including but not limited to hardware support.

8. References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1764–1772, Bejing, China. PMLR.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Jocelyn Huang, Oleksii Kuchaiev, Patrick O'Neill, Vitaly Lavrukhin, Jason Li, Adriana Flores, Georg Kucsko, and Boris Ginsburg. 2020. Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition. *arXiv preprint arXiv:2005.04290*.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. 2019. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*.

Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. 2017. Transfer learning for speech recognition on a budget. *arXiv preprint arXiv:1706.00290*.

Peter Mihajlik, Andras Balog, Tekla Etelka Graczi, Anna Kohari, Balázs Tarján, and Katalin Mady. 2022a. BEA-base: A benchmark for ASR of spontaneous Hungarian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1970–1977, Marseille, France. European Language Resources Association.

Péter Mihajlik, András Balog, Tekla Etelka Grászi, Anna Kohári, Balázs Tarján, and Katalin Mády. 2022b. Bea-base: A benchmark for asr of spontaneous hungarian. *arXiv preprint arXiv:2202.00601*.

NVIDIA. 2024. Nemo toolkit core adapters. Accessed: 2024-04-07.

Bethan Thomas, Samuel Kessler, and Salah Karout. 2022. Efficient adapter transfer of self-supervised speech models for automatic speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7102–7106. IEEE.