POLISH-JAPANESE ACADEMY OF INFORMATION TECHNOLOGY

Synthetic Boosted Automatic Speech Recognition

Rolczyński Rafał

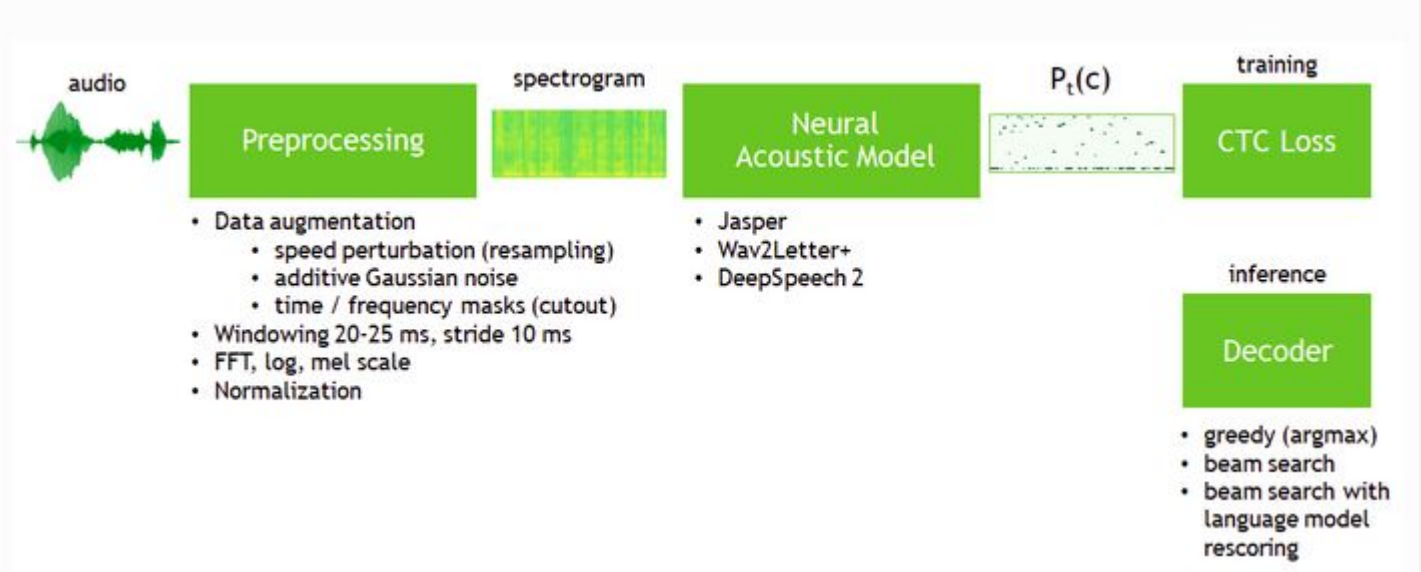
Supervisor: Prof. dr hab. Krzysztof Marasek

Agenda

1. Introduction
2. Data
3. Model
4. Optimization
5. Experiments
6. Evaluation
7. Conclusions

Introduction - Problem Definition

A deep neural network solves two closely related tasks. It learns to recognize phonemes and to formulate grammar rules at the same time. In fact, a model is able to parallel and accurate build both of them, when a training corpus is large enough.



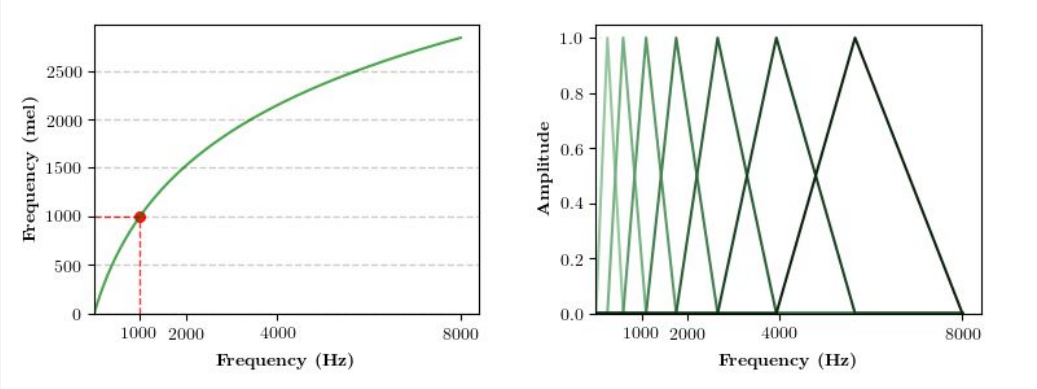
Source: https://nvidia.github.io/OpenSeq2Seq

Introduction - Goal

Use synthetic data to enrich more profound the language model

Data - Audio Representation

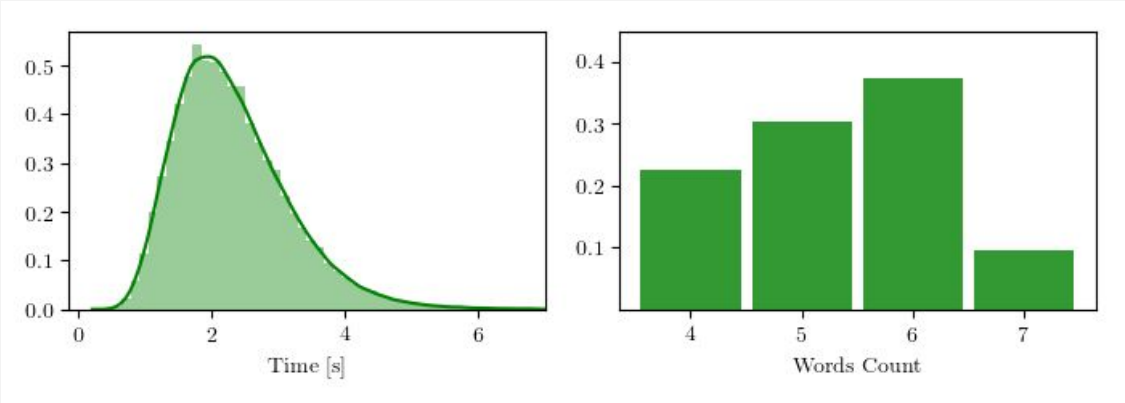
Audio samples have the single-channel (mono) with the sampling rate of 16 kHz and the depth of 16 bits. Input data the to the model is the audio signal presented in the form of Mel-scale log lter banks.



Pre-emphasis	Winlen (ms)	Winstep (ms)	Winfunc	NFFT	Filterbanks
0.95	25	10	hamming	512	80

Data - Dataset Construction

Our corpus is composed of the Clarin-PL and the Jurisdic datasets (both are well documented), and contain a variety of recording sources, including the read speech, the spontaneous dictation, and the phone calls

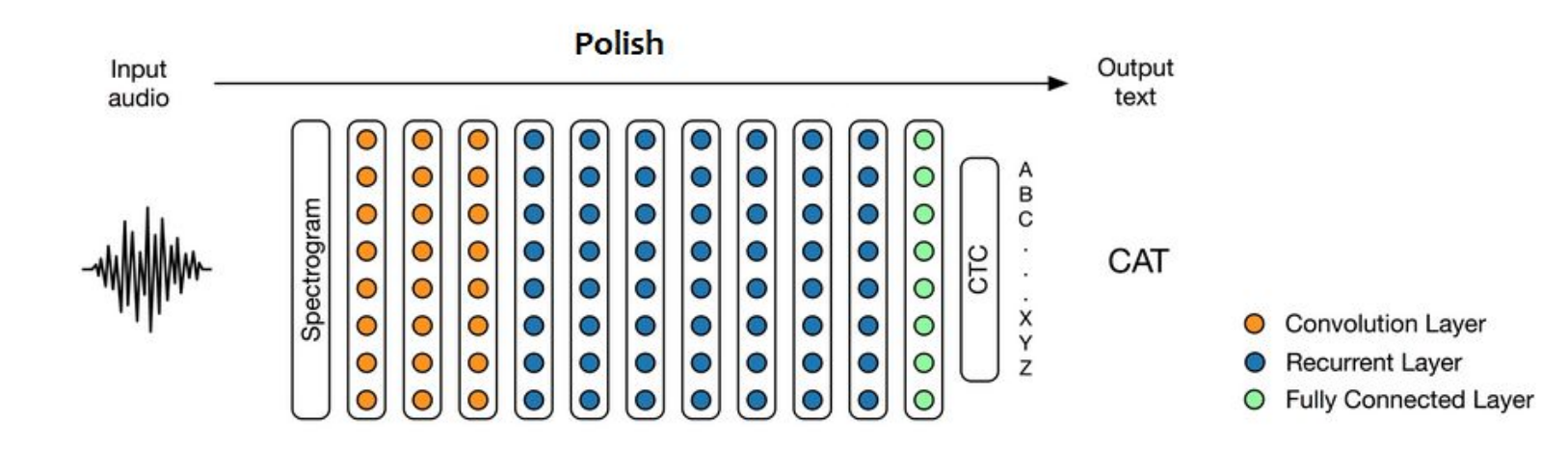


The data comes from a variety of sources is highly non-uniform, both in terms of audio data and transcriptions. Therefore, we perform the segmentation of samples into single words, and then combine into utterances from 4 to 7 words.

Corpus	Size (hours)	Samples (k)
Clarin-PL (train)	34	41
Jurisdic	351	578

Model - Base Model

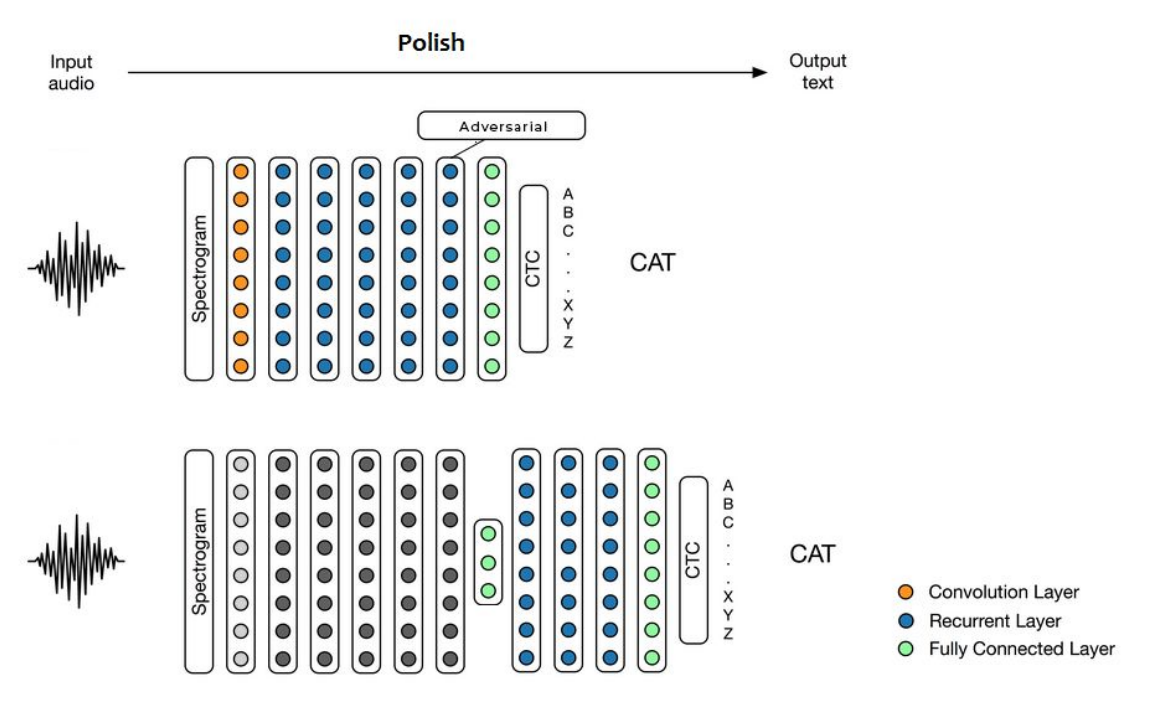
The base model architecture is built upon the Deep Speech 2 model. The model contains a convolutional layer and then several recurrent layers. At the end of the model, there is an output layer that, which for each time step returns the probability distributions over each letter in the alphabet.



Source: https://svail.github.io/mandarin

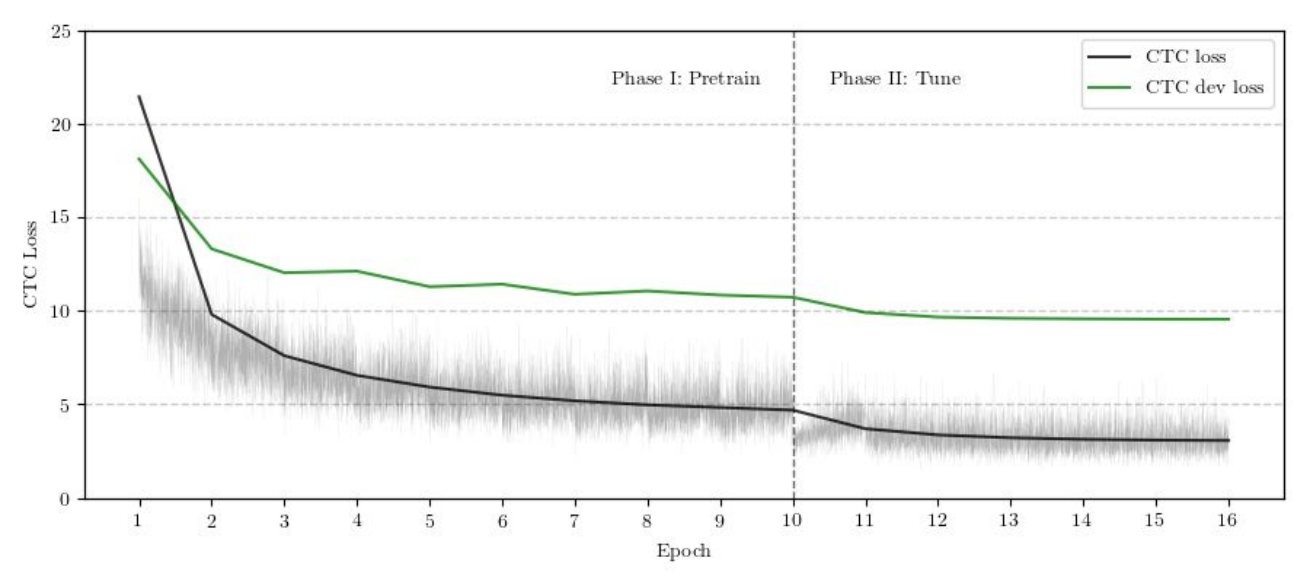
Model - Synthetic Boosted Model

The training is divided into two stages. Firstly, we train the **Phoneme Model** on a dataset with the ratio 1 to 1, authentic to synthetic data. Then, we train the **Synthetic Language Model**, which processes several times larger synthetically augmented training corpus. The Phoneme Model is frozen (weights are not updated) to protect the ability of actual phoneme recognition, therefore it plays a feature extractor role exclusively at this stage. Finally, the purpose of the Synthetic Language Model is to *boost* the representation created by the frozen Phoneme Model, and return the correct prediction.



Optimization

The optimization aims to find the model parameters, for which we minimize the loss function. For this purpose, we calculate the gradient and iteratively update the parameters according to the optimization method, which in our case is the Adaptive Moment Estimation (Adam) method. The process is divided into two phases. In the first *pretrain* phase, the learning rate is constant, and the relatively large learning which rate aims to overcome a high plateau, The second ***tune*** phase aims to fine-tune parameters in the discovered stable parameters region.



Experiments - Entire Dataset

Phoneme Model: The first layer is the two-dimensional Convolutional Layer, where the receptive field size is 15x41 (time-frequency), and the number of channels equals 32. Then, there are 5 bidirectional recurrent lstm layers with the width of 650. At the end of the model, there is a linear layer with an activation function softmax, which is removed after the training. The objective function is the weighted sum of the CTC Loss and the Adversarial Loss (theirs weights are equal).

Synthetic Language Model: The first layer is the projection layer with the width of 36, and the activation function clipped rectified-linear with the boundary value of 20. Then there are 3 bidirectional recurrent LSTM layers with the width of 650.
At the end of the model, there is a linear layer with the activation function softmax.

WER	CER	CTC Loss	Auth. Audio	Syn. Audio	Model
16,60	3,39	4,80	385		Base
15,21	3,12	4,25	385	300	Phoneme
14,33	2,78	3,99	385	1000	Boosted

Experiments - Limited Dataset

Phoneme Model: The first layer is the two-dimensional Convolutional Layer, where the receptive field size is 15x41 (time-frequency), and the number of channels equals 32. Then, there are 5 bidirectional recurrent lstm layers with the width of 650. At the end of the model, there is a linear layer with an activation function softmax, which is removed after the training. The objective function is the weighted sum of the CTC Loss and the Adversarial Loss (theirs weights are equal).

Synthetic Language Model: The first layer is the projection layer with the width of 36, and the activation function clipped rectified-linear with the boundary value of 20. Then there are 3 bidirectional recurrent LSTM layers with the width of 650.
At the end of the model, there is a linear layer with the activation function softmax.

WER	CER	CTC Loss	Auth. Audio	Syn. Audio	Model
21,69	4,22	5,71	34		Base
21,37	4,20	5,92	34		Base+
21,13	4,20	5,66	34	34	Phoneme
19,81	3,96	5,67	34	1000	Boosted
16,60	3,39	4,80	385		Base

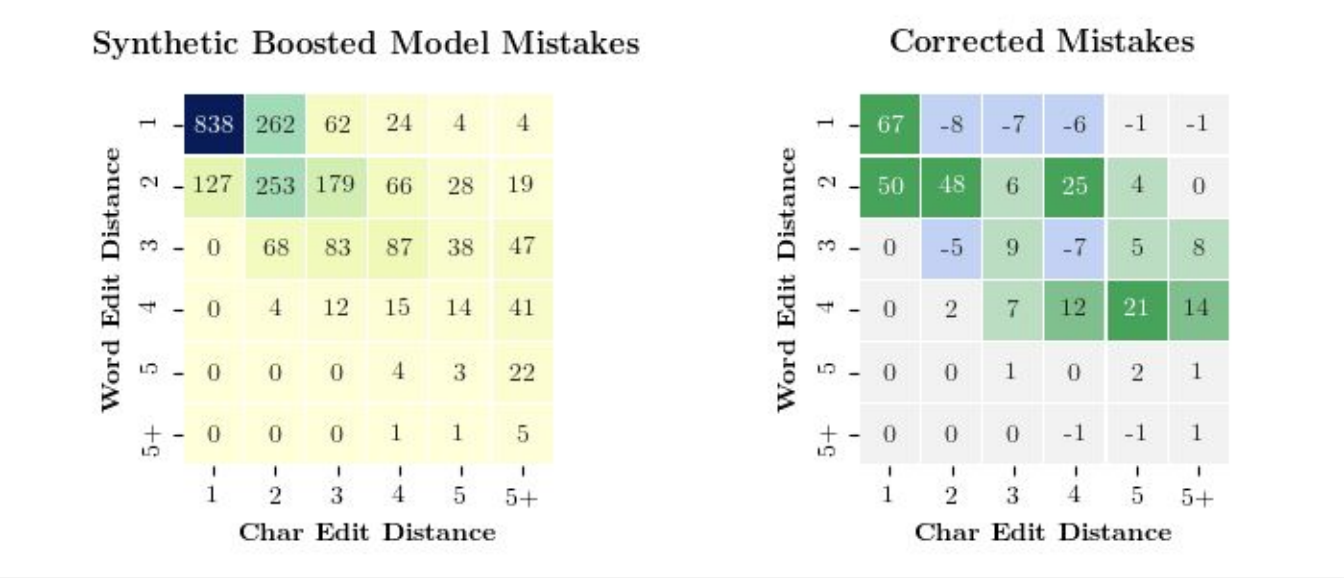
Evaluation - Results

We do the evaluation on the hold-out Clarin-test dataset, which has independent speakers and unique transcriptions. The Synthetic Boosted Model (the model named Boosted in the table) trained on the entire synthetically augmented dataset achieves more than **12.5%** absolute better the WER score, which equals **14,53%**.

WER-test	CER-test	WER-dev	CER-dev	Auth. Audio	Syn. Audio	Model
16,64	3,34	16,60	3,39	385		Base
14,53	2,98	14,33	2,78	385	1000	Boosted

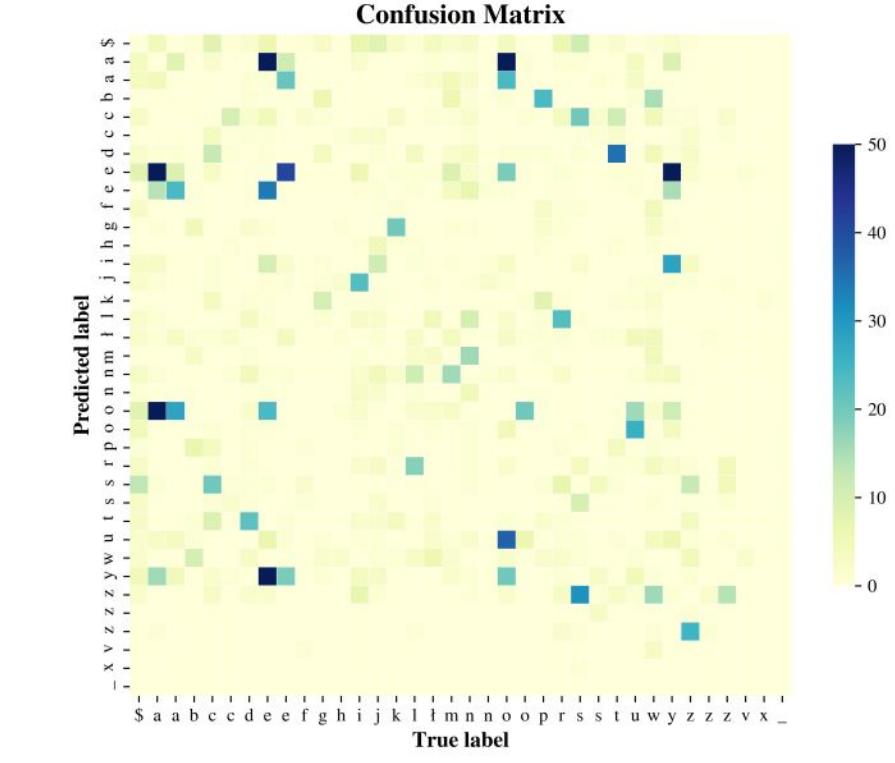
Evaluation - Error Analysis

On the left is the matrix grouping mistakes of the Synthetic Boosted Model by the Char Edit Distance and the Word Edit Distance. We can see that more than 800 samples have a single spelling mistake. On the right side, we see the corrected mistakes with respect to the Base Model. The most difficult to correct mistakes are on the out of a diagonal, which have high values of the Char Edit Distance. We can observe that the Synthetic Language Model also corrects mistakes that have high Char Edit Distance values.



Evaluation - Error Analysis

The largest group of misspellings is caused by the wrong character prediction. The overall pattern is that vowels and consonants are grouped. There are various concrete patterns which can be easily interpreted. One of them is that the chars frequently occurred at the end of words (a, e, y) are often each other misspelled. The letters which sound similar, but depend on the context are written differently, such as t and d, s and z, or p and b. Finally, the different grammar rules, for instance, the u and ó, the j and i are hard to be correctly defined.



Conclusions

The presented **Synthetic Boosted Model** has successfully improved the performance of the automatic speech recognition end-to-end neural network model. The vast amount of synthetic data can be used to increase the language information supplied to the model. The improvement is achieved thanks to the new model architecture, the new objective function, and the new training policy.

Future Works

New Model Architecture

New Adversarial Component

Specialized Domain Adaptation

Thank you

Bibliography

1. A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, Deep speech: Scaling up end-to-end speech recognition. 1412.5567, 2014.
2. D. Amodei, R. Anubhai, and E. B. et al., Deep speech 2: End-to-end speech recognition in english and mandarin. 1512.02595, 2015.
3. G. Zweig, C. Yu, J. Droppo, and A. Stolcke, Advances in All-Neural Speech Recognition, sep 2016.
4. J. Kim, M. El-Khamy, and J. Lee, Residual LSTM: Design of a Deep Recurrent Architecture for Distant Speech Recognition, jan 2017.
5. K. J. Han, A. Chandrashekaran, J. Kim, and I. Lane, The CAPIO 2017 Conversational Speech Recognition System, dec 2017.
6. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, SpecAugment:A Simple Data Augmentation Method for Automatic Speech Recognition, apr 2019
7. A. Zeyer, K. Irie, R. Schlüter, and H. Ney, Improved training of end-to-end attention models for speech recognition, may 2018.