# Joint Audio/Text Training for Transformer Rescorer of Streaming Speech Recognition

**Suyoun Kim** **Ke Li** **Lucas Kabela**
**Rongqing Huang** **Jiedan Zhu** **Ozlem Kalinli** **Duc Le**
Meta, USA
suyounkim@meta.com

## Abstract

Recently, there has been an increasing interest in two-pass streaming end-to-end speech recognition (ASR) that incorporates a 2nd-pass rescoring model on top of the conventional 1st-pass streaming ASR model to improve recognition accuracy while keeping latency low. One of the latest 2nd-pass rescoring model, Transformer Rescorer, takes the n-best initial outputs and audio embeddings from the 1st-pass model, and then choose the best output by rescoring the n-best initial outputs. However, training this Transformer Rescorer requires expensive paired audio-text training data because the model uses audio embeddings as input. In this work, we present our Joint Audio/Text training method for Transformer Rescorer, to leverage unpaired text-only data which is relatively cheaper than paired audio-text data. We evaluate Transformer Rescorer with our Joint Audio/Text training on Librispeech dataset as well as our large-scale in-house dataset and show that our training method can improve word error rate (WER) significantly compared to standard Transformer Rescorer without requiring any extra model parameters or latency.

## 1 Introduction

Streaming end-to-end automatic speech recognition (ASR) models aim to transcribe the user's voice with minimal latency and have been widely used in numerous interactive ASR applications that support direct user interaction in real-time. Unlike non-streaming end-to-end ASR (Chorowski et al., 2015; Chan et al., 2016; Bahdanau et al., 2016; Kim et al., 2017; Chiu et al., 2018), streaming end-to-end ASR, such as RNN-T (Graves, 2012a; Prabhavalkar et al., 2017; Battenberg et al., 2017; He et al., 2019; Li et al., 2019), are limited to use short audio context or not use future context to satisfy low latency constraints and suffer from higher word error rates (WER).

To address this issue of streaming ASR, a Two-Pass architectures has been recently proposed to improve WER while keeping latency low (Sainath et al., 2019; Li et al., 2020; Xu et al., 2022). The main idea of the Two-Pass architectures is to use a non-streaming model, so-called *Rescorer*, (2nd-pass) on top of the conventional streaming model (1st-pass) to re-score and choose the best output among the initial n-best outputs generated from 1st-pass model. Specifically, the latest Transformer-based (Vaswani et al., 2017) Rescorer (Li et al., 2020) has been shown promising results in both accuracy and latency improvement. The Rescorer can improve WER and keep low latency because 1) it does not need to perform expensive beam search decoding process which is done already with 1st-pass model, and 2) it exploits full context of audio embeddings generated from the 1st-pass model. While there are performance advantages to exploit full context of audio embeddings as input, training Transformer Rescorer needs expensive paired audio/text data similar to ASR training and thus cannot leverage unpaired text-only data.

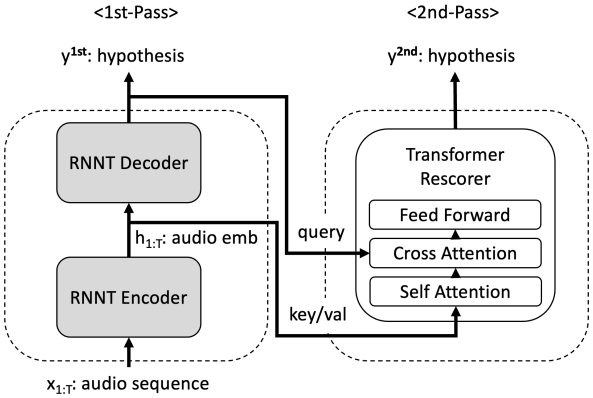


Figure 1: The two-pass system consists of streaming 1st pass RNN-T model and non-streaming 2nd pass Transformer Rescorer.

Many previous studies have investigated approaches to leverage unpaired text-only data including fusion with an external neural network language model (NNLM) (Gulcehre et al., 2015; Kan-

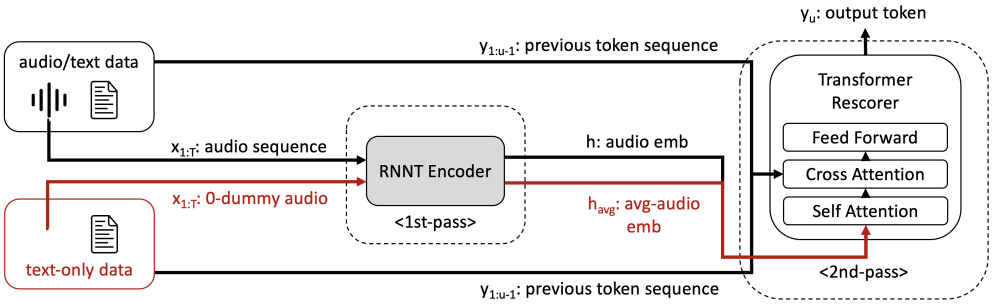


Figure 2: Our Joint Audio/Text training for Transformer Rescorer in ASR

nan et al., 2018; Sriram et al., 2017; Shan et al., 2019; Toshniwal et al., 2018; Kannan et al., 2018; McDermott et al., 2019; Variani et al., 2020; Weinstein et al., 2020; Kim et al., 2021b), however, these techniques require extra model parameters and latency. More recently, new approaches to use text-only data without requiring extra model parameters and latency by multi-task learning for LAS (Chan et al., 2016) have been proposed (Sainath et al., 2020; Wang et al., 2020; Tang et al., 2022).

In this work, we present our Joint Audio/Text training method for Transformer Rescorer to leverage unpaired text-only data without requiring any extra parameter or latency. Unlike previous studies (Sainath et al., 2020; Wang et al., 2020), our method does not need to use Text-To-Speech (TTS) or prior information (domain ID) or tuning parameter, and it is based on Transformer decoder. We evaluate our Transformer Rescorer with our Joint Audio/Text training method on Librispeech dataset as well as large-scale in-house dataset and show that our model significantly outperforms over Transformer Rescorer with standard training method WER without requiring extra training parameters or computational latency.

## 2 Transformer Rescorer

The architecture of the two-pass system (Sainath et al., 2019) consists of streaming 1st pass RNN-T model (Graves, 2012a; Shi et al., 2021) and non-streaming 2nd pass Transformer Rescorer (Li et al., 2020) is illustrated in Figure 1. The standard way to train the two-pass system is two-step:

1. Given paired input acoustic frames, $\boldsymbol{x} = (x_1, \cdots, x_T)$, and corresponding output text, $\boldsymbol{y} = (y_1, \cdots, x_U)$, the Encoder of RNN-T generates high-level audio embeddings, $\boldsymbol{h} = (h_1, \cdots, h_T)$. The Decoder of RNN-T takes $\boldsymbol{h}$ and $\boldsymbol{y}$ and generate initial output $\hat{\boldsymbol{y}}^{\boldsymbol{1st}}$ in a streaming fashion. The RNN-T model is trained to maximize $P(\boldsymbol{y} = \hat{\boldsymbol{y}}|\boldsymbol{x})$ (Graves, 2012b) first. The parameters of RNN-T are then fixed.

2. Then, Transformer Rescorer is trained with audio embeddings $\boldsymbol{h}$ generated from the fixed encoder of RNN-T and true transcription text $\boldsymbol{y}$ to maximize $\sum_u logP(y_u|\boldsymbol{h}, y_{<u})$. (Vaswani et al., 2017).

In step 2, Transformer Rescorer is based on Transformer Decoder setup (Vaswani et al., 2017), and the true transcription, $\boldsymbol{y}$, is the query and audio embeddings, $\boldsymbol{h}$, is the key in cross attention.

During inference, the 1st-pass RNN-T model generates n-best initial hypotheses $\hat{\boldsymbol{y}}^{1st} = (\hat{\boldsymbol{y_1}}^{1st}, \cdots, \hat{\boldsymbol{y_n}}^{1st})$ with standard beam search process. Then, Transformer Rescorer takes the n-best initial hypotheses as well as the full audio embeddings $\boldsymbol{h_{1:T}}$ from RNN-T model and computes the log probability score for each hypothesis (re-score) and the final best hypothesis $\hat{\boldsymbol{y}}^{\boldsymbol{2nd}}$ is generated.

## 3 Joint Audio/Text Training

Our Joint Audio/Text Training method aims to leverage large amount of unpaired text-only data without requiring extra model parameter or latency. To do so, we allow the model to be trained on either 1) paired audio/text data where both modality inputs are available, or 2) unpaired text data where only text inputs are available. Our training method is also two-step as described in 2.

In step 2, for training on the unpaired text-only example, we use the estimated averaged audio embeddings, $\boldsymbol{h_{avg}}$, of paired audio/text data in the training set. This averaged audio embeddings, $\boldsymbol{h_{avg}}$, can be simply obtained by passing a 0-dummy audio sequence to the well-trained Encoder of RNN-T from the step 1. Our Transformer Rescorer takes this averaged audio embeddings

instead as the keys and values in cross attention. Unlike (Sainath et al., 2020; Wang et al., 2020), our approach is based on Transformer architecture and it does not need to change the original objective function, nor does it require a tuning parameter for multiple losses or any prior knowledge of inputs. Figure 2 illustrates our proposed Joint Audio/Text training method for Transformer Rescorer. Note that the inference process is the same as general rescorer as described in 2.

## 4 Experiments

### 4.1 Data

**Librispeech** We first evaluate our approach on the Librispeech English corpus (Panayotov et al., 2015) which is publicly available. The training data contains 960 hours of labeled speech and an additional text-only corpus containing 810M words. We apply spectrum augmentation (Park et al., 2019) and speed perturbation. The Librispeech unpaired text-only corpus contains 810M words and 40M samples, which is almost 27 times bigger than the paired audio/text data. We will discuss the effect of data mixing ratio of text-only data in Section 5.1

**Large-Scale In-house Voice Command** We also evaluate our approach on our large-scale in-house English dataset as well. Our in-house training dataset has two sources: 20K hours of publicly shared video data and 20K hours of voice assistant domain data. All videos and audios are completely de-identified. We augment the training data with speed perturbation, simulated room impulse response, and background noise, resulting in 145K hours. Unlike Librispeech, we have 4 times smaller size of unpaired text-only corpus then paired audio-text data. The unpaired text-only data contains 1M samples of in-domain (VA) text-only data and 25M samples of general-domain text-only data. Our evaluation data, VA has 44.2K de-identified short-form utterances in the voice assistant domain, collected by a third-party data vendor.

### 4.2 Model

For the 1st pass model, we use an RNN-T model architecture which is widely used in streaming ASR (Graves, 2012b; Graves et al., 2013). The RNN-T consists of Encoder and Decoder. Our Encoder has 20 emformer layers (Shi et al., 2021). We extract 80-channel filterbanks features and convert to 320-dimensional inputs with a stride of 4. The context size is 160ms for streaming restriction. The Predictor in Decoder consists of 3 LSTM layers. The Joiner in Decoder uses 5k word-piece as our targets. Our 1st-pass RNN-T model has 79M parameters.

For the 2nd pass model, Transformer Rescorer, we use 2 layers of conventional Transformer decoder (Vaswani et al., 2017) contains both the self-attention and the cross-attention. The attention dimension is 1024 and feed forward dimension is 4K, and use 8 multi-headed attention. Transformer Rescorer takes the hypothesis from the RNN-T Decoder as a query and the 1024-dimenional audio embeddings from the RNN-T Encoder as a key/value in the cross-attention. Our 2nd-pass Transformer Rescorer model has 44M parameters. Note that our Joint Audio/Text training does not require any extra model parameter. The architectures of the 1st pass RNN-T model and the 2nd pass Transformer Rescorer are illustrated in Figure 1.

The RNN-T and Transformer rescorer are trained in two steps as described in 2 and 3. For Librispeech experiments, we trained RNN-T for 120 epochs with ADAM optimizer and a base learning rate of 0.001. For large-scale in-house experiments, we trained 15 epochs with same scheduler until full convergence.

During inference, we used the standard beam search with a beam size of 10 and generated 10-best hypotheses. As described in 2, Once initial decoding is done from RNN-T model, 10-best hypotheses and full context audio embeddings from RNN-T are passed to Transformer rescorer in parallel. For Librispeech experiments, we did not use any external neural language model. For large-scale in-house experiments, we used a small neural language model(2.5M parameters) for decoding with shallow fusion (Kim et al., 2021a) and ILME (Meng et al., 2021) to compare our method with the best baseline system.

We evaluated the speech engine perceived latency (SPL) for the latency analysis. The SPL measures the time from the end of the user's utterance until the ASR engine completes the result. We evaluated the SPL for three models: 1) 1st pass RNNT baseline without 2nd pass Transformer Rescorer, 2) Baseline with 2nd pass Transformer Rescorer, and 3) Baseline with 2nd pass Transformer Rescorer trained with our proposed Joint Audio/Text training method. The averaged SPLs (measured in ms) were 633.0, 636.0, 636.0, for models 1), 2) and 3), respectively. Overall, although the use of the 2nd

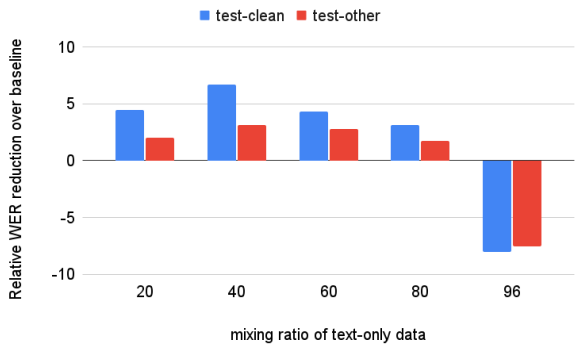


Figure 3: WER improvement of Rescorer with different `mixing ratio` on Librispeech.

pass model can increase SPL, our proposed Joint Audio/Text training method itself does not increase the latency at all because it does not require any model architecture changes.

## 5 Results

### 5.1 Effect of Mixing Ratio

Similar to previous study (Wang et al., 2020), we observed that text-only data mixing ratio is crucial to succeed with Joint Audio/Text training for Transformer rescorer as well. Figure 3 shows the relative WER reduction of the Rescorer on Librispeech text-clean/test-other with different `mixing ratio`, defined as follows:

$$\texttt{mixing ratio} = \frac{\text{\# of text-only}}{\text{\# of (text-only + audio-text)}}$$

Note that when we use the entire text-only data from the Librispeech provided, the mixing ratio was 96%. The baseline in Figure 3 was 0% mixing ratio which means that we used Rescorer trained only on paired audio/text data. We observed that text-only data of 40% `mixing ratio` performed best, and adding more than 80% of text-only data even performed worse than the baseline Rescorer. Based on this observation, we used 40% and 50% `mixing ratio` for the experiments in 5.2 and 5.3, respectively.

### 5.2 Results on Librispeech

Table 1 shows WER results on Librispeech test-clean/test-other evaluation set with Baseline (BS), Baseline with Transformer Rescorer (BS + RS), and the baseline with Transformer Rescorer with our Joint audio/text training (BS + RS + Our Joint A/T). As we discussed in Section 5.1, we used 40% mixing ratio to obtain the best results. We observed that 8.9% and 9.4% WER relative improvements by using standard Transformer Rescorer(RS), and 14.9% and 12.2% WER relative improvement by using RS trained with our Joint A/T.

### 5.3 Large-Scale In-house dataset

Table 2 shows WER results on 44.2K in-house evaluation set with Baseline (BS), Baseline with Transformer Rescorer (BS + RS), and the baseline with the Transformer Rescorer with our Joint audio/text training (BS + RS + Our Joint A/T). As previously described in Section 4.1, the text-only data in test domain was only 1% and the text-only data in general domain was only 19% among the entire training data. In this experiment, we over-sampled in-domain text-only data to 20% and out-of-domain text-only data to 30%, thus we use 50% `mixing-ratio`. We observed that 4.9% WER relative improvements by using standard rescorer, and 7.8% WER relative improvement by using rescorer trained with our Joint A/T. Surprisingly, our approach was still effective with our strong baseline which was trained on 145K hours. We also observed that using over-sampled duplicated in-domain text-only data is more effective rather than using unique out-of-domain text-only data.

## 6 Conclusions

We have introduced Joint Audio/Text training method for Transformer Rescorer of the streaming two-pass end-to-end ASR. Unlike standard training method for Transformer Rescorer, our method can leverage unpaired text-only data and consequently improves recognition accuracy without requiring extra model parameters or computational latency. We evaluated our approach on the Librispeech dataset as well as large-scale in-house dataset and showed that Transformer Rescorer with our proposed method obtained 3% - 7% relative improvement in WER compared to the standard Transformer Rescorer model.

| Models | test-clean | test-other |
|---|---|---|
| BS | 3.59 | 9.10 |
| BS + RS | 3.27 | 8.25 |
| BS + RS + Our Joint A/T | **3.06** | **7.99** |

Table 1: Comparison of WER on Librispeech with the baseline (BS), BS with the standard rescoring model (BS + RS), and BS with RS trained by our Joint Audio/Text (BS + RS + Our Joint A/T).

| Models | VA |
|---|---|
| BS | 8.32 |
| BS + RS | 7.91 |
| BS + RS + Our Joint A/T | **7.67** |

Table 2: Comparison of WER on large-scale in-house dataset. VA is our in-house 20K hours of voice assistant domain data (described in Section 4.1.)

## Limitations

As with the majority of studies, this study has potential limitation. The primary limitation is that the benefit of our training approach that leverages unpaired text-only data may be diminished when paired audio/text training data is abundant and they are in same domain as test domain.

## References


Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2016. End-to-end attention-based large vocabulary speech recognition. In *ICASSP*. IEEE.

Eric Battenberg, Jitong Chen, Rewon Child, Adam Coates, Yashesh Gaur Yi Li, Hairong Liu, Sanjeev Satheesh, Anuroop Sriram, and Zhenyao Zhu. 2017. Exploring neural transducers for end-to-end speech recognition. In *ASRU*, pages 206–213. IEEE.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP*. IEEE.

C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani. 2018. State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In *ICASSP*.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *NeurIPS*.

A. Graves. 2012a. Sequence transduction with recurrent neural networks. In *ICML Representation Learning Workshop*.

Alex Graves. 2012b. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *ICASSP*, pages 6645–6649. IEEE.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, et al. 2019. Streaming end-to-end speech recognition for mobile devices. In *ICASSP*, pages 6381–6385. IEEE.

Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. 2018. An analysis of incorporating an external language model into a sequence-to-sequence model. In *ICASSP*, pages 1–5828. IEEE.

S. Kim, Y. Shangguan, J. Mahadeokar, A. Bruguier, C. Fuegen, M. L. Seltzer, and D. Le. 2021a. Improved Neural Language Model Fusion for Streaming Recurrent Neural Network Transducer. In *Proc. ICASSP*.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *ICASSP*. IEEE.

Suyoun Kim, Yuan Shangguan, Jay Mahadeokar, Antoine Bruguier, Christian Fuegen, Michael L Seltzer, and Duc Le. 2021b. Improved neural language model fusion for streaming recurrent neural network transducer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7333–7337. IEEE.

Jinyu Li, Rui Zhao, Hu Hu, and Yifan Gong. 2019. Improving rnn transducer modeling for end-to-end speech recognition. In *ASRU*, pages 114–121. IEEE.

Wei Li, James Qin, Chung-Cheng Chiu, Ruoming Pang, and Yanzhang He. 2020. Parallel rescoring with transformer for streaming on-device speech recognition. *arXiv preprint arXiv:2008.13093*.

Eric McDermott, Hasim Sak, and Ehsan Variani. 2019. A density ratio approach to language model fusion in end-to-end automatic speech recognition. In *ASRU*. IEEE.

Z. Meng, S. Parthasarathy, E. Sun, Y. Gaur, N. Kanda, L. Lu, X. Chen, R. Zhao, J. Li, and Y. Gong. 2021. Internal Language Model Estimation for Domain-Adaptive End-to-End Speech Recognition. In *Proc. SLT*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *ICASSP*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

R. Prabhavalkar, K. Rao, T. Sainath, B. Li, L. Johnson, and N. Jaitly. 2017. A Comparison of Sequence-to-Sequence Models for Speech Recognition. In *Interspeech*, pages 939–943.

Tara N Sainath, Ruoming Pang, David Rybach, Yanzhang He, Rohit Prabhavalkar, Wei Li, Mirkó Visontai, Qiao Liang, Trevor Strohman, Yonghui Wu, et al. 2019. Two-pass end-to-end speech recognition. *arXiv preprint arXiv:1908.10992*.

Tara N Sainath, Ruoming Pang, Ron J Weiss, Yanzhang He, Chung-cheng Chiu, and Trevor Strohman. 2020. An attention-based joint acoustic and text on-device end-to-end model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7039–7043. IEEE.

Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, Dong Yu, and Lei Xie. 2019. Component fusion: Learning replaceable language model component for end-to-end speech recognition system. In *ICASSP*, pages 5361–5635. IEEE.

Y. Shi, Y. Wang, C. Wu, C. Yeh, J. Chan, F. Zhang, D. Le, and M. L. Seltzer. 2021. Emformer: Efficient Memory Transformer Based Acoustic Model For Low Latency Streaming Speech Recognition. In *Proc. ICASSP*.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2017. Cold fusion: Training seq2seq models together with language models. *arXiv preprint arXiv:1708.06426*.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, et al. 2022. Unified speech-text pre-training for speech translation and recognition. *arXiv preprint arXiv:2204.05409*.

Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, and Karen Livescu. 2018. A comparison of techniques for language model integration in encoder-decoder speech recognition. In *SLT*, pages 369–375. IEEE.

Ehsan Variani, David Rybach, Cyril Allauzen, and Michael Riley. 2020. Hybrid autoregressive transducer (hat). In *ICASSP*, pages 6139–6143. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Peidong Wang, Tara N Sainath, and Ron J Weiss. 2020. Multitask training with text data for end-to-end speech recognition. *arXiv preprint arXiv:2010.14318*.

Eugene Weinstein, James Apfel, Mohammadreza Ghodsi, Rodrigo Cabrera, and Xiaofeng Liu. 2020. Rnn-transducer with stateless prediction network. In *ICASSP*, pages 7049–7053.

Liyan Xu, Yile Gu, Jari Kolehmainen, Haidar Khan, Ankur Gandhe, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko. 2022. Rescorebert: Discriminative speech recognition rescoring with bert. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6117–6121. IEEE.