# MAX-MARGIN TRANSDUCER LOSS: IMPROVING SEQUENCE-DISCRIMINATIVE TRAINING USING A LARGE-MARGIN LEARNING STRATEGY

*Rupak Vignesh Swaminathan, Grant P. Strimel, Ariya Rastrow, Harish Mallidi, Kai Zhen, Hieu Duy Nguyen, Nathan Susanj, Athanasios Mouchtaris*

Amazon.com Inc., United States

## ABSTRACT

In this work, we propose a novel sequence-discriminative training criterion for automatic speech recognition (ASR) based on the Conformer Transducer. Inspired by the large-margin classifier framework, we separate the "good" and the "bad" hypotheses in an N-best list produced from a pre-trained transducer model by a margin ($\tau$), hence the term, Max-Margin Transducer (MMT) loss. It is observed that fine-tuning with the proposed loss achieves significant improvement over baseline transducer loss but does not outperform the state-of-the-art minimum word error rate (MWER) training. However, combining the proposed MMT loss with MWER surpasses the performance of either losses suggesting the complimentary nature of MWER and MMT losses. With the combined losses, we obtained 7.44% and 7.68% relative WER improvements on Librispeech test-clean and test-other sets, respectively, and up to 8.9% relative improvement on Multi-lingual Librispeech test sets.

***Index Terms*—** Max-margin, Conformer, minimum word error rate training, sequence discriminative criterion, end-to-end speech recognition models

## 1. INTRODUCTION

End-to-end ASR systems based on Transducers [1, 2] have gained increasing popularity over the recent years owing to their streaming capabilities and their simplicity over the traditional factored ASR systems. These benefits have led to their wide adoption in personal voice assistants and speech services [3, 4, 5, 6]. Typically, transducer model parameters are optimized using a maximum likelihood estimation criteria which aims to increase the likelihood of all possible label-to-audio alignments. This is followed by a sequence-discriminative fine-tuning stage [7, 8, 9] which ties the training objective to the final metric, for example, the expected word error rate (WER). Although sequence-discriminative training such as MWER offers improvements for transducer-based models [7, 8], there is still opportunity for improving discrimination of incorrect hypotheses from the promising ones in terms of their alignment scores that affect the beam pruning.

In this work, we adopt a large-margin style training, most popularly utilized in Support Vector Machines (SVM) [10], and combine it with sequence-discriminative fine-tuning by using a margin term to separate the "good" and "bad" hypotheses present in the N-best list. Specifically, we place a margin on the transducer alignment scores obtained from hypotheses that correspond to ground truth token sequences (positive examples) and other hypotheses in the N-best list that have word errors (negative examples). We formulate this MMT loss similar to the MWER loss, a type of Minimum Bayes Risk (MBR) optimization, and fine-tune network parameters such that the emitted hypotheses have a good separation between the positive and negative examples. This allows us to ultimately combine the MMT and MWER loss in the most efficient manner. Consequently, we observe a significant improvement when incorporating both MMT and MWER losses compared to either of them in isolation, offering up to 7.68% relative improvements on Librispeech test sets and up to 8.9% relative improvements on Multi-lingual Librispeech over the baseline transducer loss.

## 2. RELATION TO PRIOR WORK

As a type of sequence-discriminative training, our work is related to a general class of training methodology known as Minimum Bayes Risk (MBR). Traditional hybrid ASR system pre-trained with frame level cross entropy perform fine-tuning of models with variants of MBR loss such as minimum phone error (MPE) and state-level MBR (sMBR) [11, 12, 13]. These objectives tie the training criterion with the final task-specific model performance criterion. MBR variants have also been successfully applied for end-to-end ASR models such as Reccurent Neural Aligner (RNA) [14], attention-based encoder-decoder (AED) [15] and RNN-T [7, 8]. In [7], MBR is implemented by computing the expected edit distance between a ground truth label sequence and an on-the-fly generated N-best list with a limited beam size and number of alignments. [8] proposed an efficient minimum WER training by performing offline decoding, computing the alignment scores for each hypotheses in the N-best list and combining them with the expected edit distance scores for MWER train-

ing, overcoming the limitations in [7].

There have been efforts to incorporate the large-margin technique in sequence-discriminative training. One of the earliest works was [16] in which the authors modified sequence-discriminative criteria such as the MPE [11] and maximum mutual information (MMI) [17, 18], to incorporate a margin term in sequence training for Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) systems. In the context of end-to-end ASR systems, [19, 20] introduced the concept of margin for attention-based models. In [19], the authors proposed a version of the large margin loss with only 1-best as a replacement for MWER loss in listen, attend, and spell (LAS) [21] models. It is shown that MWER still generalizes well across test sets, but large-margin loss can acheive similar performance on certain test sets with only 1-best. In [20], a novel technique known as promising accurate prefix boosting is introduced that maximizes the margin between decoded prefixes and the original ground truth prefix. The technique performs slightly better than MWER but comes at the expense of increased computational cost during the training time.

Our work is the first attempt at incorporating large-margin training for transducer models to the best of our knowledge. We utilize multiple hypotheses and utterance-level scores from the forward-backward algorithm [1] in construction of the MMT loss, thereby capturing all possible alignment paths for hypotheses in the N-best list. Furthermore, our work is also the first to combine MWER and MMT losses. We formulate our MMT loss in such a way that it re-uses significant portions of the computations used for MWER loss, allowing us to combine the two losses in the most efficient manner.

## 3. MAX-MARGIN TRANSDUCER

In this section, we first present an overview of our baseline systems, specifically, the Transducer architecture, its training method, and MWER fine-tuning, after which the Max-Margin Transducer is discussed in detail.

### 3.1. Conformer Transducer: An overview

The Transducer architecture [1, 2] consists of an encoder or a transcription network that consumes audio features and produces a high-level representation, a prediction network that takes previously emitted non-blank token as input, and a joint network that combines both the acoustic and text representations to produce a time-token grid representation through which the sum of all possible alignments $P(y^*|x)$ are maximized.

Maximizing the alignment likelihood is equivalent to minimizing the negative log-likelihood of $P(y^*|x)$. Hence, the transducer loss is defined as,

$$\mathcal{L}_{transducer} = -ln\ P(y^*|x) \tag{1}$$

### 3.2. Minimum Word Error Rate (MWER) Fine-tuning

We compare our proposed technique with MWER fine-tuning proposed in [8]. In this implementation, the N-best lists are first used to compute the word errors $R(y_i, y^*)$ where $y_i$'s are the hypotheses and $y^*$ is the ground truth label sequence, then the sum of scores of all possible alignments for each hypothesis in the N-best list is computed. Finally, the alignments scores and word errors are combined as follows (Equation 2).

$$\mathcal{L}_{mwer} = \sum_{y_i \in nbest(x)} \text{softmax}(ln\ P(y_i|x))R(y_i, y^*) \tag{2}$$

### 3.3. Max-Margin Transducer Loss

We propose a novel sequence-discriminative criteria where the decoded utterances ($y_i$) are fed back into the prediction network to generate all the possible alignments given the audio frames ($x_t$) and obtain an utterance level score for each of the hypothesis as shown in Figure 1. We can introduce a margin term $\tau$, to act on the utterance level scores ($ln\ P(y_i|x)$) as the separation factor to keep the positive (ground truth) hypothesis present in the N-best list and the negative hypotheses (utterances with errors) at least $\tau$ units apart. Here, $y^*$ indicates the most probable N-best hypothesis (positive example) that has an exact match with the ground truth and $y_i$ are the other hypotheses with errors in them (negative examples). However, this formulation will lead to stability issues and difficulty in choosing the margin term $\tau$ as the absolute utterance level scores (log-likelihoods) are data dependent. To overcome this, we convert the utterance level alignment scores, into a probability distribution by taking a softmax (referred as $S_{y_i}$).

$$\text{MaxMargin}(y^*, y_i) = \max(0, \tau - (S_{y^*} - S_{y_i})) \tag{3}$$

This way, the margin term $\tau$ is a tunable parameter in the range [0,1]. We set this MaxMargin error to 0 when a $y_i$ from N-best list is same as the ground truth $y^*$. We can then use this MaxMargin error in the formulation below in order to compute the loss and back-propagate the gradients.

$$\mathcal{L}_{mmt} = \sum_{y_i \in nbest(x)} \text{softmax}(ln\ P(y_i|x))\ \text{MaxMargin}(y^*, y_i) \tag{4}$$

An advantage of the proposed method is the fast and efficient computation of the margin loss using the forward-backward algorithm outlined in [1]. This computation gives us utterance probabilities which can be computed only once if we need to combine the MMT loss with the MWER loss as shown below.

$$\mathcal{L}_{combined} = \mathcal{L}_{mwer} + \lambda\mathcal{L}_{mmt} \tag{5}$$

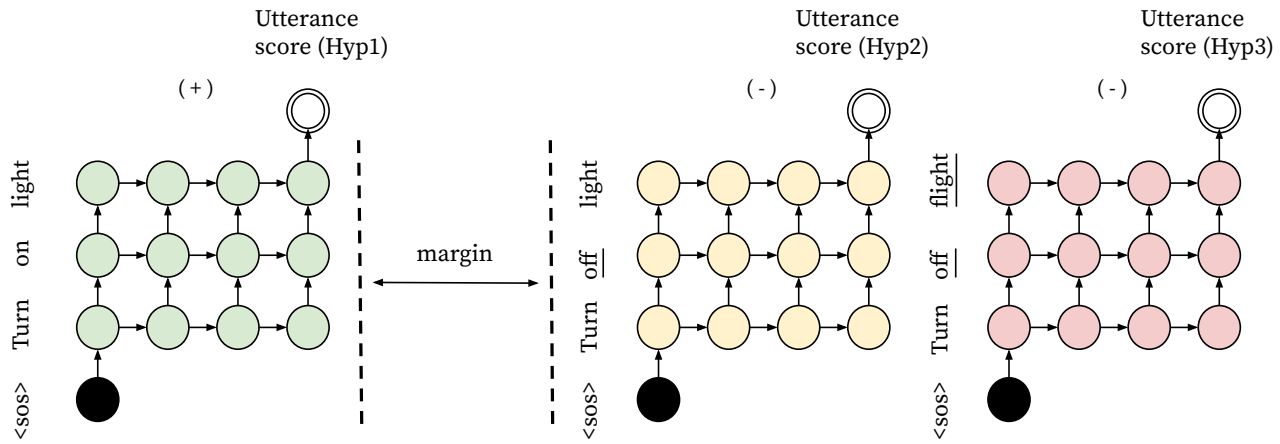


**Fig. 1**. 3-best hypotheses are shown where the ground truth sequence "Turn on light" in the N-best list is chosen as the positive example. These 3-best hypotheses are fed back into the prediction network after which the joint network lattice is computed to obtain the utterance scores (sum of all possible alignment paths) for each N-best hypothesis. In this example, the pairs (Hyp1, Hyp2) and (Hyp1, Hyp3) would contribute to the margin loss if their utterance scores are within the given margin.

Another major advantage of the proposed loss is that we consider all possible alignment paths of the N-best hypotheses in the computation of the MMT loss. Hence, error-prone utterances that are close to the ground truth in the time-token space will get pushed further away leading to a better hypotheses surviving beam pruning during inference.

## 4. EXPERIMENTAL SETUP

### 4.1. Datasets

We use Librispeech [22] and Multi-lingual Librispeech (MLS) [23] datasets to benchmark our proposed loss against the baseline transducer loss and MWER loss. In Librispeech experiments, the model is trained on the full 960 hour train partition and evaluated on the standard dev-clean, dev-other, test-clean, and test-other evaluation sets. Checkpoint selection is done using the dev-other partition. In the MLS experiments, we first train a multi-lingual model which is followed by a locale-specific finetuning to adapt the initial model to the individual locales from the MLS dataset[1].

### 4.2. Model Architecture

In our experiments, we use a Conformer Transducer [24], whose audio encoder consists of stacks of 16 Conformer blocks. Each Conformer block consists of two feed-forward layers, a multi-head attention layer, and a convolutional module. The input features are 64-dimensional LFBE coefficients, extracted with a 25ms window size and a 10ms hop size. The extracted features are then fed to the sub-sampling blocks that consist of two layers of 2D CNN with filters of 128 channels, kernel size of 3, and stride of 2, making the feature frame rate 40ms. The conformer blocks use multi-head attention with 4 heads and each head with a dimension of 64. The feed-forward hidden unit dimension is 1024. The prediction network is a single layer LSTM with 640 units and an output projection layer of 512 dimensions. The output tokens modeled by the transducer are 2.5K word pieces which include the blank symbol $\phi$. The number of trainable parameters in this model is 30M.

### 4.3. Training scheme

We use Adam optimizer [25] ($\beta_1$=0.9, $\beta_2$=0.98, $\epsilon$=1e-9) with the same learning rate schedule described in [24]. Based on the dev-other WER, we select the checkpoint after 200K steps. The MWER/MMT models are initialized with this checkpoint and are fine-tuned for 10K steps with a learning rate of 5e-5. We train this model using 24 Tesla V100 GPUs. It takes 50 hours for the main training, and 10 hours for MWER/MMT fine-tuning. We use SpecAugment (mask parameter F = 27, time masks = 20, maximum time-mask ratio pS = 0.05), for the transducer training. We do not use any pre-training or language models in these experiments. Our baseline Librispeech model achieves WER comparable to the results reported in [26][2]. For the MWER/MMT finetuning, we turn off SpecAugment as this generates N-best lists that are closer to the ones produced by the model in the inference environment. The beam size is set to 6 and beam width is set to 8 during both N-best list generation and evaluation. We add an auxiliary transducer loss (weight=1e-3) during sequence finetuning to stabilize the training [27]. We also artificially

[1]We excluded pl-PL in our setup due to a high OOV rate

[2]The MLS baseline numbers reported in this paper are higher than those in [23] due to a smaller model size, choice of output tokens and loss function.

**Table 1**. Results on development and test sets of Multi-lingual Librispeech for various loss types.

| Loss Type | en-US | | de-DE | | es-ES | | fr-FR | | it-IT | | nl-NL | | pt-BR | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | dev | test | dev | test | dev | test | dev | test | dev | test | dev | test | dev | test |
| $\mathcal{L}_{transducer}$ | 8.81 | 10.50 | 9.98 | 11.12 | 8.38 | 9.74 | 13.70 | 13.33 | 19.40 | 17.60 | 20.46 | 21.52 | 21.52 | 17.72 |
| $\mathcal{L}_{mwer}$ | 8.49 | 10.12 | 9.57 | 10.65 | 7.75 | 9.02 | 12.98 | 12.65 | 18.23 | 16.52 | 19.01 | 20.01 | 20.31 | 16.74 |
| $\mathcal{L}_{mmt}$ | 8.52 | 10.13 | 9.59 | 10.70 | 7.83 | 9.12 | 13.05 | 12.73 | 18.44 | 16.72 | 19.13 | 20.16 | 20.59 | 16.99 |
| $\mathcal{L}_{combined}$ | **8.36** | **9.97** | **9.44** | **10.51** | **7.63** | **8.88** | **12.78** | **12.46** | **17.95** | **16.28** | **18.74** | **19.79** | **20.01** | **16.46** |

add the ground truth sequence to the N-best list generated for MWER/MMT computation if they are not present already as this helps in the presence of "positive" examples for MMT.

## 5. RESULTS AND DISCUSSION

In Table 2, we report the results of our proposed approach for Librispeech test sets. We tune the margin parameter with $\tau$=(0.1, 0.3, 0.5, 0.7, 0.9) and report the best tuning results ($\tau$=0.3). For the $\mathcal{L}_{combined}$ experiments, we tune the value of the MMT loss weight $\lambda$=(1.0, 1e-1, 1e-2, 1e-3), however, we find that the $\lambda$ value of 1.0 yields the best results. This equal weighting suggests that the overall loss gives equal importance to word errors and margin errors which reflects in the WER as well. It can be observed that MWER offers significant improvements in the range of 5.67% - 6.84% relative WER reduction across evaluation sets of Librispeech (3.36→3.13 on test-clean, 7.55→7.12 on test-other). It is also worth noting that MMT loss in isolation offers commendable gains although not as much as MWER. However, the combination of the losses $\mathcal{L}_{combined}$ outperform both MWER and MMT losses in isolation offering up to 7.68% relative WER reduction (3.36→3.11 on test-clean, 7.55→6.97 on test-other), implying that the losses are largely complimentary and offer additive gains to the model performance.

**Table 2**. Absolute WER numbers on development and test sets of Librispeech are reported for various loss types.

| Loss Type | dev-clean | dev-other | test-clean | test-other |
|---|---|---|---|---|
| $\mathcal{L}_{transducer}$ | 3.01 | 7.40 | 3.36 | 7.55 |
| $\mathcal{L}_{mwer}$ | 2.82 | 6.98 | 3.13 | 7.12 |
| $\mathcal{L}_{mmt}$ | 2.9 | 6.92 | 3.16 | 7.26 |
| $\mathcal{L}_{combined}$ | **2.77** | **6.86** | **3.11** | **6.97** |

We applied the best margin term and weighting hyper-parameters ($\tau$=0.3, $\lambda$=1.0) to the MLS experiments and the results reported in Table 1 show a trend similar to Librispeech results. MWER models for various locales show relative WER improvements in the range of 3.5% (en-US) to 7.5% (es-ES). The MMT counterparts also show significant improvements but falling short of the performance of MWER models. But when the losses are combined with equal weighting one can observe an improvement (4.9% for en-US and 8.9% for es-ES in relative WER improvements) that is greater than either of the losses.

To observe the effect of the margin term $\tau$ on WER, we sweep the values from 0.1 to 0.9 while keeping the weighting constant ($\lambda$=1.0). In Table 3, we can notice that for the lowest $\tau$, the overall loss does not accumulate many margin errors and the result is quite close to the MWER loss in isolation. As we increase the $\tau$, we find a sweet spot at 0.3 where we achieve the best WER, beyond which the model's hypotheses start to degrade due to very aggressive $\tau$.

**Table 3**. Effect of margin parameter $\tau$ on WER for Librispeech test sets shown below (Loss Type $\mathcal{L}_{combined}$).

| $\tau$ | dev-clean | dev-other | test-clean | test-other |
|---|---|---|---|---|
| 0.1 | 2.81 | 6.95 | 3.13 | 7.05 |
| **0.3** | **2.77** | **6.86** | **3.11** | **6.97** |
| 0.5 | 2.86 | 6.98 | 3.15 | 7.11 |
| 0.7 | 3.05 | 7.46 | 3.39 | 7.66 |
| 0.9 | 3.30 | 7.75 | 3.65 | 7.87 |

## 6. CONCLUSION

In this work, we introduced the Max-Margin Transducer loss, a novel sequence-discriminative training strategy for Conformer Transducer. Specifically, we leverage the MMT formulation on utterance scores of the hypotheses from the N-best list to separate a correct hypothesis from confusing hypotheses if the scores are within a margin $\tau$, a tunable hyper-parameter. We show that the proposed MMT loss performs as good as MWER, and the combined MMT+MWER loss significantly outperforms both MMT and MWER in isolation offering up to 7.68% relative WER improvement on Librispeech test sets and up to 8.9% on Multilingual Librispeech test sets.

## 7. REFERENCES

[1] Alex Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[2] Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, et al., "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.

[3] Mahaveer Jain, Kjell Schubert, Jay Mahadeokar, Ching-Feng Yeh, Kaustubh Kalgaonkar, Anuroop Sriram, Christian Fuegen, and Michael L Seltzer, "Rnn-t for latency controlled asr with improved beam search," *arXiv preprint arXiv:1911.01629*, 2019.

[4] Suhaila M Shakiah, Rupak Vignesh Swaminathan, Hieu Duy Nguyen, Raviteja Chinta, Tariq Afzal, Nathan Susanj, Athanasios Mouchtaris, Grant P Strimel, and Ariya Rastrow, "Accelerator-aware training for transducer-based speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 100–107.

[5] Martin Radfar, Rohit Barnwal, Rupak Vignesh Swaminathan, Feng-Ju Chang, Grant P Strimel, Nathan Susanj, and Athanasios Mouchtaris, "Convrnn-t: Convolutional augmented recurrent neural network transducers for streaming speech recognition," 2022.

[6] Rupak Vignesh Swaminathan, Brian King, Grant P Strimel, Jasha Droppo, and Athanasios Mouchtaris, "Codert: Distilling encoder representations with co-learning for transducer-based speech recognition," *arXiv preprint arXiv:2106.07734*, 2021.

[7] Chao Weng, Chengzhu Yu, Jia Cui, Chunlei Zhang, and Dong Yu, "Minimum bayes risk training of rnn-transducer for end-to-end speech recognition," *arXiv preprint arXiv:1911.12487*, 2019.

[8] Jinxi Guo, Gautam Tiwari, Jasha Droppo, Maarten Van Segbroeck, Che-Wei Huang, Andreas Stolcke, and Roland Maas, "Efficient minimum word error rate training of rnn-transducer for end-to-end speech recognition," *arXiv preprint arXiv:2007.13802*, 2020.

[9] Cal Peyser, Tara N Sainath, and Golan Pundak, "Improving proper noun recognition in end-to-end asr by customization of the mwer loss criterion," in *ICASSP 2020*. IEEE, 2020, pp. 7789–7793.

[10] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] Daniel Povey and Philip C Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2002, vol. 1, pp. I–105.

[12] Daniel Povey, *Discriminative training for large vocabulary speech recognition*, Ph.D. thesis, University of Cambridge, 2005.

[13] Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks.," in *Interspeech*, 2013, vol. 2013, pp. 2345–2349.

[14] Hasim Sak, Matt Shannon, Kanishka Rao, and Françoise Beaufays, "Recurrent neural aligner: An encoder-decoder neural network model for sequence to sequence mapping.," in *Interspeech*, 2017, vol. 8, pp. 1298–1302.

[15] Rohit Prabhavalkar, Tara N Sainath, Yonghui Wu, Patrick Nguyen, Zhifeng Chen, Chung-Cheng Chiu, and Anjuli Kannan, "Minimum word error rate training for attention-based sequence-to-sequence models," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 4839–4843.

[16] Georg Heigold, Thomas Deselaers, Ralf Schlüter, and Hermann Ney, "Modified mmi/mpe: A direct evaluation of the margin in speech recognition," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 384–391.

[17] Lalit Bahl, Peter Brown, Peter De Souza, and Robert Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *ICASSP 1986*. IEEE, 1986, vol. 11, pp. 49–52.

[18] Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, and Karthik Visweswariah, "Boosted mmi for model and feature-space discriminative training," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4057–4060.

[19] Peidong Wang, Jia Cui, Chao Weng, and Dong Yu, "Large margin training for attention based end-to-end speech recognition.," in *INTERSPEECH*, 2019, pp. 246–250.

[20] Murali Karthick Baskar, Lukáš Burget, Shinji Watanabe, Martin Karafiát, Takaaki Hori, and Jan Honza Černockỳ, "Promising accurate prefix boosting for sequence-to-sequence asr," in *ICASSP 2019*. IEEE, 2019, pp. 5646–5650.

[21] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[22] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE ICASSP*. IEEE, 2015, pp. 5206–5210.

[23] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

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

[25] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2017.

[26] Martin Radfar, Paulina Lyskawa, Brandon Trujillo, Yi Xie, Kai Zhen, Jahn Heymann, Denis Filimonov, Grant Strimel, Nathan Susanj, and Athanasios Mouchtaris, "Conmer: Streaming conformer without self-attention for interactive voice assistants," in *Interspeech 2023*, 2023.

[27] Zhiyun Lu, Yanwei Pan, Thibault Doutre, Parisa Haghani, Liangliang Cao, Rohit Prabhavalkar, Chao Zhang, and Trevor Strohman, "Input length matters: Improving rnn-t and mwer training for long-form telephony speech recognition," 2021.