Improving Sinhala Speech Recognition Through e2e LF-MMI Model

Buddhi Gamage, Randil Pushpananda, Thilini Nadungodage, Ruvan Weerasinghe
Language Technology Research Laboratory (LTRL)
University of Colombo School of Computing, Sri Lanka
{bud, rpn, hnd, arw}@ucsc.cmb.ac.lk

Abstract

Automatic speech recognition (ASR) has experienced several paradigm shifts over the years from template-based approaches and statistical modeling to the popular GMM-HMM approach and then to deep learning hybrid model DNN-HMM. The latest shift is to end-to-end (e2e) DNN architecture. We present a study to build an e2e ASR system using state-of-the-art deep learning models to verify the applicability of e2e ASR models for the highly inflected and yet low-resource Sinhala language. We experimented on e2e Lattice-Free Maximum Mutual Information (e2e LF-MMI) model with the baseline statistical models with 40 hours of training data to evaluate. We used the same corpus for creating language models and lexicon in our previous study, which resulted in the best accuracy for the Sinhala language. We were able to achieve a Word-error-rate (WER) of 28.55% for Sinhala, only slightly worse than the existing best hybrid model. Our model, however, is more context-independent and faster for Sinhala speech recognition and so more suitable for general purpose speech-to-text translation.

1 Introduction

There are two main architectures in training the ASR system. They are Statistical ASR architecture and End-to-End(e2e) Deep Neural architecture. Statistical ASR was the state of the art for many years, however after the year 2015, researchers tend to move towards e2e ASR systems due to the higher results. The main difference between these two architectures is the number of models needed to create, in training the ASR system. In Statistical ASR, it needs 3 types of models and they are acoustic models, pronunciation models and the language model. But in e2e ASR, it will compress those 3 models into a single Deep-Neural-Network (DNN) (Wang et al., 2019).

So many research have been done using e2e architecture for creating ASR systems for different languages since this is a new trending area of Natural-Language-Processing (NLP) and speech recognition. And there are previous researches conducted for English speech recognition that show better results when using e2e architecture than the traditional statistical approach (Park et al., 2019).

Currently, there are previous and ongoing researches on building ASR systems for Sinhala language using statistical ASR architecture, Gaussian mixture model with Hidden Markov model (GMM-HMM) based models and Hybrid Deep-Neural-Network with Hidden Markov model (DNN-HMM) based models (Gamage et al., 2020; Karunathilaka et al., 2020). E2e architecture would be a new approach for Sinhala ASR and it will help to improve the available resources as well. Especially, e2e architecture opens the doorway to transfer learning, which is the new trending for low resource speech recognition, to improve the accuracy. These domains for speech recognition were popular in later 2019 among the researchers and it is essential to have models trained in e2e architecture in Sinhala language to get a generic idea when accessing these domains (Stoian et al., 2020).

In this paper, we present a study on e2e DNN architecture based ASR systems for Sinhala speech recognition using e2e LF-MMI model. The performances of the e2e model will be evaluated and compared with the statistical models such as GMM-HMM, DNN-HMM and combinational models of SGMM-DNN.

The paper is organized as follows. Section 2 presents the related studies, Section 3 describes the methodology, data preparation and

Proceedings of the 18th International Conference on Natural Language Processing, pages 213–219
Silchar, India. December 16 - 19, 2021.

implementation in greater detail. Section 4 describes the results and evaluation. Section 5 presents the conclusions and the future work.

2 Literature review

More than 30 years later also this methodology still predominates in ASR. Nowadays, most practical speech recognition systems are based on the statistical approach (Wang et al., 2019).

With the improvement of deep learning, DNN is introduced for creating acoustic models. The role of DNN is to calculate the posterior probability of the HMM state, by replacing the traditional GMM observation probability. So DNN-HMM models become the state-of-the-art ASR model by achieving better results than GMM-HMM models (Wang et al., 2019). The training process and decoding process of the HMM-based model determines whether it faces the following difficulties in actual use.

- The training process is very complicated and it is difficult to perform global optimization.
- When constructing HMM based models, they made an assumption that the 3 models are independent from each other. This simplify the model creation but this is not an actual match (Wang et al., 2019).

Due to the above-mentioned shortcomings or anomalies in the HMM-based models, more research was carried out in the e2e architecture with the trending of deep learning. The end-to-end model is a system that directly maps input audio sequence to sequence of words or other graphemes (Rao et al., 2017). So direct mapping of utterances to character sequence is conducted where no intermediate states like calculating posteriors in the output (Wang et al., 2019).

Data alignment is the major problem in both HMM based and e2e models but e2e models require soft alignments where HMM based models use forced alignments. However, main problem in e2e architecture is that it requires a large amount of speech data to achieve higher accuracy in recognition (Wang et al., 2019). Till year 2018 low resource speech recognition systems are never used e2e architecture because this architecture is used in Large Vocabulary Continuous Speech Recognitio (LVCSR). (Povey et al., 2016) paper shows that, abundance of training data make the system lagged comparable to hybrid DNN systems when trained on smaller training sets.

Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages research conducted by Microsoft Corporation in India created a challenge by giving 50 hours of transcribed speech in each Tamil, Telugu and Gujarati which are three main indian languages and asked participants to build ASR systems restricted to this dataset. Evaluation carried out on a blind test set (Srivastava et al., 2018). The ISI-Billa system presented to address the above challenge (Billa, 2018) was an EESEN based end-to-end multilingual LSTM network trained using the Connectionist Temporal Classification (CTC) training criterion . This is the first time to use an e2e model for south asian languages. Both monolingual and multilingual systems trained using this model and it outperformed the baseline models in all three languages (Srivastava et al., 2018).

After year 2019, transfer learning and unsupervised learning techniques have become famous in the speech recognition domain. One of the major improvements was to tackle the low resource problems researchers are introducing in these domains (Bataev et al., 2018). In the Investigation of transfer learning for ASR using LF-MMI trained neural networks paper they used weight transfer and multitask learning transfer learning techniques to address the low resource problem (Ghahremani et al., 2017). Meta Learning For End-to-end Low-resource Speech Recognition paper shows that using multilingual CTC models can be used to improve the accuracy of the ASR using Meta learning techniques (Hsu et al., 2020). In wav2vec: Unsupervised Pre-training for Speech Recognition paper used a model is trained using the Auto Segmentation Criterion. Untranscribed web audio for low resource speech recognition paper has introduced semi-supervised training is done by using lattice-free maximum mutual information (LF-MMI) to untranscribed data (Carmantini et al., 2019).

3 Approach

3.1 Data Preparation

Data preparation is the most important task in the ASR pipeline because to have a reliable ASR, it highly depends on the consistency and integrity of the data preparation step (Gamage et al., 2020). We used Kaldi toolkit in the study and data preparation is done accordingly. Since the training scripts of e2e LF-MMI models does not allow segment file which has the length of each utterance in a single audio file, we had to split the audio recored in praat to have a single audio wave file to a single utterance in training models.

3.1.1 Dataset

Here we have used the collected recordings from Language Technology Research Laboratory (LTRL) of University of Colombo School of Computing (UCSC) which has 40 hours of training data which have been gathered using Praat and Redstart tools. Training the models involves a total data set from 113 native speakers where 79 are female, and 34 are males. The training data set has audio recordings from 67 females and 27 males speakers, and the total utterances are 17848 sentences, which is 25h of speech data. As the validation data set,2002 speech utterances from 8 females and 3 male speakers are taken for fine-tuning the models. Testing the models involve a data set from 4 female speakers and 4 male speakers where they utter 80 speech sentences altogether. Training has done in 16kHz sample rate and refer (Gamage et al., 2020) for more details. The overall details about the data sets are given in table 1.

Dataset	Male	Female	Utterances
Train	27	67	17848
Dev	3	8	2002
Test	4	4	80

Table 1: Details of train,validation and test data sets

3.1.2 Lexicon

Lexicon has the mapping of words with the relevant spoken phone sequences and it is a major part of the pronunciation model in a statistical ASR system (Gamage et al., 2020). For creating lexicon, "Sinhala G2P Conversion" (Nadungodage et al., 2018) and "Subasa Transliterator" tools are used and please refer (Gamage et al., 2020) for more details.

3.1.3 Corpus

We used 3 corpora namely, UCSC Novel Corpus (90000 unique sentences), Chatbot Corpus (388 unique sentences) and the corpus created using active learning method (20000 unique sentences) to create the corpus for the study. By combining the above corpora, a new corpus is created to generate ngram language models. Summary of the corpus is available in table 2. SRILM (Stolcke, 2002) toolkit and KenLM (Heafield, 2011) toolkit are used to create the n-gram language models. After calculating perplexities for testing dataset we selected a 4-gram language model through the study. Details of the Language Models are represented in table 3.

Vocabulary Size	243339
Total number of Sentences	119621
Total number of words	1194940

Table 2: Corpus Statistics

Language Model	Perplexity
Witten-Bell 3grams	9.393376
Witten-Bell 4grams	8.108833

Table 3: Perplexities of created Language Models

3.2 Baseline models

Mainly, 4 baseline models are considered in the study excluding monophone and triphone models as done for Sinhala language mentioned in (Gamage et al., 2020). Those baseline models are,

- Subspace Gaussian mixture model with MMI (SGMM+MMI)
- Hybrid System (Dan's DNN)
- Hybrid System (Keral's DNN)
- Combination SGMM + Dan's DNN

Detailed descriptions of creating baseline models are presented in (Gamage et al., 2020).Usage of Combinational Acoustic

Models(DNN-HMM and SGMM) and Identifying the Impact of Language Models in Sinhala Speech Recognition (Gamage et al., 2020) paper uses 30 hours of training data. Additionally we have used 10 more hours with the same dataset.

Mel Frequency Cepstral Coefficient (MFCC) feature extraction is done using 13 MFCC with downsampling and zero order coefficient as it is the standard measurement, and features are extracted every 10ms with the 25ms Hamming window (Povey et al., 2011). Followed by, we trained models with monophone training, triphone training with delta feature computation, Linear Discriminant Analysis (LDA) with Maximum Likelihood Linear Transform(MLLT) and Speaker Adaptive Training (SAT). Alignments of LDA+MLLT+SAT model are used to train SGMM+MMI models and Hybrid models. DNN models are influenced by Keral's recipe and Dan's recipe mentioned in Kadli (Povey et al., 2011). Parameters for above models are mentioned in (Gamage et al., 2020), in detail. Results obtained in baseline models are represented in table 4.

3.3 Proposed E2e Lattice-Free Maximum Mutual Information (LF-MMI) model

We used the default Neural Network (NN) architecture to train WSJ dataset mentioned in training recipes. We used Factored Time Delay Neural Networks (TDNNf) according to the standard Kaldi WSJ recipe. This neural network has 13 TDNNf layers and a rank reduction layer. The number of units in the TDDNf consists of 1024 and 128 bottleneck units. The default hyperparameters of the standard recipe were used with the number of epochs 10, 30 and 50. (Hadian et al., 2018).

We chose the phone based training to create the e2e models. Architecture used to create e2e LF-MMI models are represented in figure 1.

Unlike in baseline models, 40-dimensional MFCC features are extracted from 25ms frames every 10ms because it is the default used in WSJ recipe (Hadian et al., 2018). Zero mean and unit variance normalization techniques are used per-speaker basis and no other feature normalization or feature transform is used. Unlike in baseline models, we do not perform re-alignments during the training here.

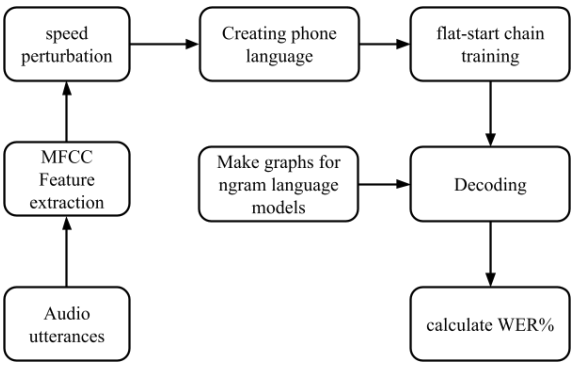

Figure 1: e2e LF-MMI model Architecture

Data is augmented with 2-fold speed perturbation in all the experiments because it modifies the length of each utterance to the nearest of the distinct lengths (Hadian et al., 2018). Otherwise, we can pad each utterance with silence to reach one of the distinct lengths.

Unlike in statistical ASR , e2e ASR decodes the utterance into character sequence. So we need a phone language for the denominator graph (Povey et al., 2016) to decode utterances. Then we start training the models in Kaldi toolkit with NN settings mentioned above. A different lang directory which contains the information of n-gram language models, can be used with a wordlist and language model of our choice to train the models, as long as phones.txt is compatible. The mkgraph.sh script helps to train the e2e models using such language models.

4 Results and Evaluation

The server used for training of all deep neural architectures and the decoding of the models contains 4 GPUs - GeForce RTX 2080 Ti of 10.8GB each. All 4 GPUs have been used when training, Thereby accelerating the deep learning training process by leveraging CUDA. The performance of e2e Sinhala ASR systems are evaluated in terms of accuracy on the recordings taken with the noise environment. This accuracy can be obtained by calculating either Word Error Rate (WER) or Sentence Error Rate (SER). WER is the number of words that are wrongly identified out of the total number of words in the audio sample used for recognition. SER is the number of sentences that are improperly identified out of the total number of sentences (Karunathi-

laka et al., 2020).The WER is widely used to discrete and compare speech recognition systems and we also used the WER throughout the study.

4.1 Baseline models

The combined SGMM+DNN statistical model is created with 40 hours of training data mentioned above. As discussed, the hybrid DNN model and SGMM+MMI models are created on top of the alignments of LDA+MMLT+SAT (tri3) triphone model. In (Gamage et al., 2020) statistical architecture achieved 31.72 %WER for only 30 hours of training data in the combined model. Table 5 represents the comparison of the results with the same architecture of previous study with 10hours more data.

Model	Test set (WERs)	Dev set (WERs)
mono	47.43	3.78
tri1	33.08	2.87
tri2	32.78	3.19
tri3	35.80	3.10
sgmm2_4	32.33	2.89
SGMM2 + MMI Training	31.72	2.87
Hybrid System (Dan's DNN)	27.79	2.29
Hybrid System (Keral's DNN)	27.95	3.96
Combination SGMM + Dan's DNN	29.00	2.44
Combination SGMM + Keral's DNN	28.40	2.95

Table 4: Results of Baseline models with 40 hours of data

We can clearly identify in table 4 that Hybrid Dan's DNN model achieved the best WER which is 27.79% and it outperformed the combined model by 0.61% less WER. With 30 hours of training data SGMM2+MMI model achieved 34.14% WER and within 40 hours it achieved 31.72% WER with the improvement of 2.42% less WER. And Hybrid Dan's model had 35.50% WER with 30 hours of training data but using 10 hours more data it achieved 27.79% WER with 5.71% improvement in WER. So we can clearly identify that with more data, hybrid DNN models performing well. This is a well known fact that DNN models have the higher accuracy in Speech recognition with the higher data available for training.

Model	WERs of	
	30 hour dataset	40 hour dataset
SGMM2 + MMI Training	34.14	31.72
Hybrid System (Dan's DNN)	35.50	27.79
Hybrid System (Keral's DNN)	36.33	27.95
Combination SGMM + Dan's DNN	31.72	29.00

Table 5: WER comparison of baseline models after increasing 10 more hours of auido data to the same dataset

4.2 E2e LF-MMI Models

Even though LF-MMI model is loosely coupled with HMM it can act as e2e model with monophones or full bi-phones by removing the context dependency tree alignments. In this study we use full left bi-phones so every possible pair of phones have a separate HMM model.

Table 6 represents the WERs achieved in e2e LF-MMI models and was able to get 28.55% WER with 10 epochs in Kaldi. When using GPU for training in Kaldi, it considers available 4 GPUs as a single GPU using exclusive mode and we can give higher frames when training. So for each epochs we used 3 million frames per iterations.

Ephocs	Test set (WERs)	Dev set (WERs)
10	28.55	2.27
30	32.18	2.90
50	33.27	2.02

Table 6: WER comparison of e2e LF-MMI models

4.3 Evaluation

Following 3 sentences are taken from 3 random persons. Recording of those utterances have been done in their own environment with their own equipment and they had 44.1 Hz sample rate. Hybrid Dan's model was taken as the baseline model because it had the lowest WER among other baseline models.They are compared with the accurate e2e LF-MMI model. Recorded utterances are fed into the above models and resulted outputs are represented in table 7 to 9.

Test sentence 1 in figure 7 is a news reading and (Gamage et al., 2020) paper shows

sentence	Russia's first humanoid robot arrives at International Space Station
Utterance	රුසියාව විසින් ප්‍රථම වරට හඳුන්වාදුන් හියුමනොයිඩ් රොබෝවරයා ජාත්‍යන්තර අභ්‍යවකාශ මධ්‍යස්ථානය වෙත ළඟා වී තිබෙ
Baseline	රුසියාව විසින් ප්‍රථමවරට හඳුන්වා දුන්නේ මා නොයෙක් ඔයා බෝර්යා ජාත්‍යන්තර අභ්‍යවකාශ මධ්‍යස්ථානය වෙත ළඟා වී තිබෙ
e2e LF-MMI	වාසිහු විසින් පැකුම් හා පහම හාඳ නො වීම නො යි ඔහු පෝමරය ජාත්‍යන්තර අභ්‍යවකාශ මධ්‍යස්ථානය වෙත ළඟා වී තිබෙ

Table 7: Analysis of test sentence 1

that the current Sinhala ASR system is well performed in the context of news readings and number readings. Because training transcriptions are mainly gathered in those areas. In this sentence also baseline model is well performed rather than other e2e models but "හියුමනොයිඩ් රොබෝවරයා" is not identified correctly. Those two words not being included in the lexicon and the corpus can be the reason. But in e2e LF-MMI model, we can see that it is trying to identify the word "හියුමනොයිඩ් " correctly by looking at "ම නො යි" "වීම නො යි".

sentence	The Cabinet also decided to provide President Maithripala Sirisena with his current official residence at Mahagama Sekara Mawatha, Colombo 7, after his retirement.
Utterance	ජනාධිපති මෛත්‍රිපාල සිරිසේන මහතා දැනට භාවිතාකරන කොළඹ හත මහගමසේකර මාවතේ පිහිටි නිලනිවස විශ්‍රාම ගැනීමෙන් පසුව ද ඔහුට ලබා දීමට කැබිනට් මණ්ඩලය තීරණය කරයි
Baseline	ඇස් ජනාධිපති මට පන්සල මහතා ඇයට භාවිතාකරන කොළඹ හත මහගමසේකර මාවතේ පිහිටි නිල නිවස විශ්‍රාම ගැනීමෙන් පසුව ද ඔහුට ලබා දීමට කැබිනට් මණ්ඩලය තීරණය කළහ
e2e LF-MMI	නාදපති මෛත්‍රිපාල සිරිසේන මහතා දැඩි භාවිතාකරන කොළඹ හත මහගමසේකර මාවතේ පිහිටි නිලනිවසට විශ්‍රාම ගැනීමෙන් පසුව ද ඔහුට ලබා දීමට කැවිලි බණ්ඩනය තීරණයක

Table 8: Analysis of test sentence 2

"ජනාධිපති මෛත්‍රිපාල සිරිසේන" and "කැබිනට් මණ්ඩලය" is not identified correctly in all models at to some extent. Those words are included in both lexicon and corpus but they are not identified correctly. Problem here is acoustic data is not enough to train, to have higher probability for those words because they are proper nouns. So n-gram language model dominated here to have higher probability in "ජනාධිපති මට" rather than "ජනාධිපති මෛත්‍රිපාල" because in previous study with less data those two words correctly identified in the baseline model. E2e model is trying to get the more accurate decoding. So we can identify that there is a misleading in Statistical models when using higher data. E2e models use character level decoding rather than word level decoding done in statistical architecture so that misleading is minimum in the e2e models and we can identify that in the above sentence.

sentence	I leave
Utterance	මම යනවා
Baseline	විවර
e2e LF-MMI	මම යනවා

Table 9: Analysis of test sentence 3

Test sentence 3 in figure 9 has a normal day to day talking accent. E2e LF-MMI model has correctly identify the utterance. Baseline model completely mis-identifies the utterance because it is context dependent when decoding and sentences with less number of words mostly have low accuracy because it uses word level decoding. Many other sentences had this observation so that we can identify that the e2e models are more context independent rather than statistical models even though the training data has a context dependency. To have a general ASR system, e2e techniques are more suitable.

5 Conclusion and Future Work

Even though there is a slightly better accuracy in statistical approach, using the e2e approach we created a less context dependent and faster model for Sinhala speech recognition for using general purpose speech-to-text transcription. We found out that using only statistical models (GMM+HMM, SGMM, SGMM+MMI) is not useful in further research conducted for Sinhala speech recognition that even hybrid system where DNN uses to calculate the posterior probabilities for the HMM perform far better than those traditional approaches.

Currently, we were able to achieve 28.55% WER for Sinhala e2e speech recognition using e2e LF-MMI implemented on Kaldi toolkit. This model also can be improved with fine tuning but this study is not going to fine tune and has used the basic implementations and recipes available for the WSJ dataset which have 80 hours of training data. Current domain of speech recognition moves toward addressing the low resource problem. There are large datasets available for English and France like languages with the state-of-the-art results. Common solution for addressing the low resource problem is to transfer learning from high resource language to a low resource lan-

guage. In e2e LF-MMI technique transfer learning can be done by using weight transfer and multi-task training (Ghahremani et al., 2017). So from the results of this study it will be useful to do the necessary data augmentations and choosing parameters for transfer learning techniques mentioned above.

References

Vladimir Bataev, Maxim Korenevsky, Ivan Medennikov, and Alexander Zatvornitskiy. 2018. Exploring end-to-end techniques for low-resource speech recognition. In International Conference on Speech and Computer, pages 32–41. Springer.

Jayadev Billa. 2018. Isi asr system for the low resource speech recognition challenge for indian languages. In INTERSPEECH, pages 3207–3211.

Andrea Carmantini, Peter Bell, and Steve Renals. 2019. Untranscribed web audio for low resource speech recognition. In INTERSPEECH, pages 226–230.

Buddhi Gamage, Randil Pushpananda, Ruvan Weerasinghe, and Thilini Nadungodage. 2020. Usage of combinational acoustic models (dnn-hmm and sgmm) and identifying the impact of language models in sinhala speech recognition. In 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pages 17–22. IEEE.

Pegah Ghahremani, Vimal Manohar, Hossein Hadian, Daniel Povey, and Sanjeev Khudanpur. 2017. Investigation of transfer learning for asr using lf-mmi trained neural networks. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 279–286. IEEE.

Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur. 2018. End-to-end speech recognition using lattice-free mmi. In Interspeech, pages 12–16.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In Proceedings of the sixth workshop on statistical machine translation, pages 187–197.

Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee. 2020. Meta learning for end-to-end low-resource speech recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7844–7848. IEEE.

Hirunika Karunathilaka, Viraj Welgama, Thilini Nadungodage, and Ruvan Weerasinghe. 2020. Low-resource sinhala speech recognition using deep learning. In 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pages 196–201. IEEE.

Thilini Nadungodage, Chamila Liyanage, Amathri Prerera, Randil Pushpananda, and Ruvan Weerasinghe. 2018. Sinhala g2p conversion for speech processing. In SLTU, pages 112–116.

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.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding, CONF. IEEE Signal Processing Society.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for asr based on lattice-free mmi. In Interspeech, pages 2751–2755.

Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. 2017. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 193–199. IEEE.

Brij Mohan Lal Srivastava, Sunayana Sitaram, Rupesh Kumar Mehta, Krishna Doss Mohan, Pallavi Matani, Sandeepkumar Satpal, Kalika Bali, Radhakrishnan Srikanth, and Niranjan Nayak. 2018. Interspeech 2018 low resource automatic speech recognition challenge for indian languages. In SLTU, pages 11–14.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7909–7913. IEEE.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In Seventh international conference on spoken language processing.

Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019. An overview of end-to-end automatic speech recognition. Symmetry, 11(8):1018.