Indic Languages Automatic Speech Recognition using Meta-Learning Approach

Anugunj Naman
IIIT Guwahati, India
anugunj.naman@iiitg.ac.in

Kumari Deepshikha
LOWE's, Bengaluru
kumari.deepshikha@lowes.com

Abstract

Recently Conformer-based models have shown promising leads to Automatic Speech Recognition (ASR), outperforming transformer-based networks while meta-learning has been extremely useful in modeling deep learning networks with a scarcity of abundant data. In this work, we use Conformers to model both global and local dependencies of an audio sequence in a very parameter-efficient way and meta-learn the initialization parameters from several languages during training to attain fast adaptation on the unseen target languages, using model-agnostic meta-learning algorithm (MAML). We analyse and evaluate the proposed approach for seven different Indic languages. Preliminary results showed that the proposed method, MAML-ASR, comes significantly closer to state-of-the-art monolingual Automatic Speech Recognition for all seven different Indic languages in terms of character error rate.

1 Introduction

"Ok, Google. Hi Alexa. Hey Siri." have featured an enormous boom of smart speakers in recent years, unveiling a trend towards ubiquitous and ambient computing (AI) for better daily lives. As the communication bridge between humans and machines, multilingual ASR is of central importance. India is a country with an enormous amount of languages and catering to those languages is difficult without having a large amount of label training corpora. Pretraining on other language sources as the initialization, then fine-tuning on target language is the main approach for such low-resource setting, also referred to as multilingual transfer learning pretraining (Multi-ASR) (Vu et al., 2014) (Tong et al., 2017). Multi-ASR models are designed to learn using an encoder to extract language-independent representations to build a better acoustic model from many source languages. The success of language independent features to improve ASR performance compared to monolingual training has been shown in many recent works (Dalmia et al., 2018) (Cho et al., 2018)(Tong et al., 2018). However, there performance have been lacklustre compared to model trained directly using target language, i.e., training for single language only.

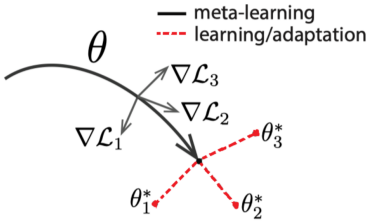

Figure 1: The MAML algorithm learns a good parameter initializer θ by training across various meta-tasks such that it can adapt quickly to new tasks.

In this paper, we follow on the concept of multilingual pretraining – Meta-learning. Meta-learning, or learning-to-learn, has recently received considerable interest within the machine learning community. The goal of meta-learning is to resolve the matter of fast adaptation on unseen data, which is aligned with our low-resource setting for different Indic languages. We use model-agnostic meta-learning algorithm (MAML) (Finn et al., 2017) in this work. As its name suggests and seen in figure 1, MAML can be applied to any neural network architecture since it only modifies the optimization process following a meta-learning training method. It doesn't introduce any additional modules like adversarial training or requires phoneme level annotation like hierarchical approaches (Hsu et al., 2019).

In recent times, the Transformer architecture based on self-attention (Zhang et al.,

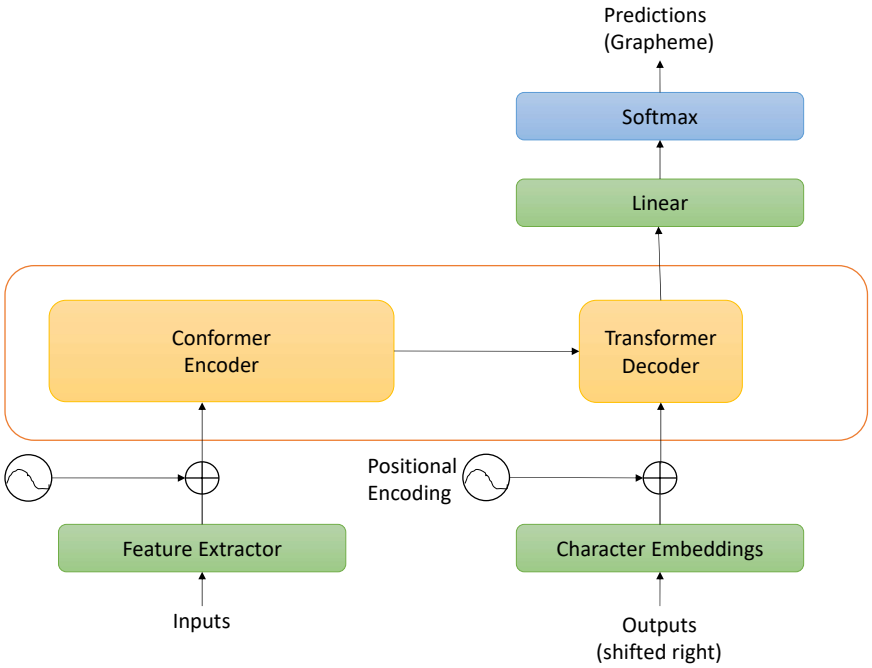

Figure 2: Transformer ASR model architecture.

2020)(Vaswani et al., 2017) has shown widespread adoption for modeling sequences due to its ability to capture long-distance interactions and the high training efficiency. Alternatively, convolutions have also been successful for speech recognition (Li et al., 2019) (Kriman et al., 2019)(Han et al., 2020)(Sainath et al., 2013)(Abdel-Hamid et al., 2014), that capture local context progressively using local receptive field layer by layer.

However, models with convolutions or self-attention each have their own limitations. While Transformers are good at modeling long-range global context, they are not very capable to extract fine-grained local feature patterns. Convolution networks, on the other hand, exploit local information and are used as the common computational block in vision. They learn shared position-based kernels over a local window which maintains translation equivariance and can capture features like edges and shapes. One limitation of using local connectivity is that you need several layers or parameters to capture global information. To tackle this issue, contemporary work ContextNet (Han et al., 2020) adopts the squeeze-and-excitation module (Hu et al., 2018) in each residual block to capture longer context. However, the model is still limited in capturing dynamic global context because it only applies a global averaging over the entire sequence.

Recently, combining convolution and self-attention has shown significant improvement in automatic speech recognition model as they can learn both position-wise local features and use content-based global interactions. We have used Conformers (Gulati et al., 2020) in this work. Conformers are the combination of self-attention and convolution sandwiched between a pair of feed-forward modules that achieves the best of both worlds i.e., self-attention learns the global interaction whilst the convolutions coherently captures the relative offset-based local correlations.

We evaluated the effectiveness of the proposed model of several Indic languages. Our experiments show that our model comes close to monolingual models.

2 Proposed Method

In this section, we present the architecture of our conformer-based speech recognition model and the proposed meta-learning method for fast adaptation to the multilingual speech recognition task.

2.1 Conformer Speech Recognition Model

As shown in Figure 2, we build our model using a Conformers to learn to predict graphemes from the

speech input. Our model extracts learnable features from audio inputs using a feature extractor module to generate input embeddings. The encoder process the input embeddings generated from the feature extractor module using conformer blocks. Mathematically, this means, for input x_i to a Conformer block i, the output z_i of the block is:

$$\begin{aligned}\tilde{x_i} &= x_i + \frac{1}{2}\text{FFN}(x_i)\\ x'_i &= \tilde{x_i} + \text{MHSA}(\tilde{x_i})\\ x''_i &= x'_i + \text{Conv}(x'_i)\\ z_i &= \text{Layernorm}(x''_i + \frac{1}{2}\text{FFN}(x''_i))\end{aligned} \tag{1}$$

where FFN refers to the Feedforward module, MHSA refers to the Multi-Head Self-Attention module, and Conv refers to the Convolution module as described in the preceding sections (Gulati et al., 2020).

Then the decoder receives the encoder outputs from conformer blocks and applies multi-head attention to its input to finally compute the logits of the outputs. To generate the probability of the outputs, we then compute the value of logits using a softmax function. We also apply a mask in the attention layer to avoid any possible information flow from future tokens. We then train our model by optimizing the next-step prediction on the previous characters and by maximizing the log probability shown below:

$$\max_{\theta} \sum_{i} \log P(y_i|z, y'_{<i}; \theta), \tag{2}$$

where z is the character inputs, y_i is the next predicted character, and $y'_{<i}$ is the ground truth of the previous characters. uring inference, we generate the output sequence using a beam-search method in an auto-regressive manner. Then we maximize the following objective function:

$$\eta \sum_{i} \log P(y_i|z, \hat{y}_{<i}; \theta) + \gamma \sqrt{wc(\hat{y}_{<i})}, \tag{3}$$

where η is the parameter to control the decoding probability from the decoder, and γ is the parameter to control the effect of the word count $wc(\hat{y}_{<i})$ as suggested in (Winata et al., 2019) and (Winata et al., 2020).

2.2 Fast Adaptation via Meta-Learning

Model-agnostic meta-learning (MAML) (Finn et al., 2017) learns to quickly adapt to a new task from a number of different tasks using a gradient descent method. In this paper, we apply MAML to effectively learn from a set of languages and quickly adapt to a new language in the few-shot setting. We denote our Conformer based ASR as f_θ parameterized by θ. Our dataset is consist a set of languages $\mathcal{A} = \{A_1, A_2, \cdots, A_n\}$, and for each language i, we split the data into A_i^{tra} and A_i^{val}, then update θ into θ' by computing gradient descent updates on A_i^{tra}:

$$\theta'_i = \theta - \alpha \nabla_\theta \mathcal{L}_{A_i^{tra}}(f_\theta), \tag{4}$$

where α is the fast adaptation learning rate. During the training, the model parameters are trained to optimize the performance of the adapted model $f(\theta'_i)$ on unseen A_i^{val}. The meta-objective is defined as follows:

$$\min_{\theta} \sum_{A_i \sim p(\mathcal{A})} \mathcal{L}_{A_i^{val}}(f_{\theta'_i}) = \sum_{A_i \sim p(\mathcal{A})} \mathcal{L}_{A_i^{val}}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{A_i^{tra}}(f_\theta)}). \tag{5}$$

where $\mathcal{L}_{A_i^{val}}(f_{\theta'_i})$ is the loss evaluated on A_i^{val}. We collect the loss $\mathcal{L}_{A_i^{val}}(f_{\theta'_i})$ from a batch of languages and perform the meta-optimization as follows:

$$\theta \leftarrow \theta - \beta \sum_{A_i \sim p(\mathcal{A})} \nabla_\theta \mathcal{L}_{A_i^{val}}(f_{\theta'_i}), \tag{6}$$

where β is the meta step size and $f_{\theta'_i}$ is the adapted network on language A_i. The meta-gradient update step is performed to achieve a good initialization for our conformer based ASR model, then we can optimize our model with few number of samples on target languages in the fine-tuning step. In this

Table 1: Statistics of Indic Language Speech Data.

Language	# Samples
Assamese **(as)**	36,000
Bengali **(be)**	232,537
Hindi **(hi)**	80,000
Marathi **(ma)**	44,500
Nepali **(ne)**	157,905
Sinhala **(sh)**	185,293
Tamil **(ta)**	62,000
Total	798,235

Table 2: Average Character Error Rate (% CER) comparison with single training.

Languages	**MAML**						**Single Training**
-	10%-shot	25%-shot	50%-shot	75%-shot	all-shot		-
Assamese	61.29	50.87	41.80	25.44	13.44	**(+1.86)**	11.58
Bengali	57.48	47.60	38.19	26.47	10.77	**(+2.04)**	8.73
Hindi	55.49	43.81	35.78	23.43	10.19	**(+2.92)**	7.27
Marathi	56.78	45.30	36.68	23.56	10.04	**(+2.91)**	7.13
Nepali	57.33	47.33	35.27	22.85	10.32	**(+3.46)**	6.86
Sinhala	54.36	45.22	35.15	24.36	11.69	**(+4.36)**	7.33
Tamil	60.38	48.70	39.89	27.41	19.74	**(+4.21)**	15.53

Table 3: Mean Human Evaluation score(0-5) for Indic Languages

Language	**MAML**		**Single Training**	
-	**Mean Correct**	**Mean Fluency**	**Mean Correct**	**Mean Fluency**
Assamese **(as)**	4.1	4.0	4.4	4.5
Bengali **(be)**	4.0	4.0	4.4	4.4
Hindi **(hi)**	4.2	4.1	4.4	4.5
Marathi **(ma)**	4.0	4.1	4.5	4.5
Nepali **(ne)**	4.1	4.0	4.6	4.5
Sinhala **(sh)**	4.1	4.1	4.5	4.4
Tamil **(ta)**	3.9	4.1	4.2	4.2

work, we use first order approximation MAML (Gu et al., 2018) and (Finn et al., 2018), thus Equation 6 is further rewritten as:

$$\theta \leftarrow \theta - \beta \sum_{A_i \sim p(\mathcal{A})} \nabla_{\theta_i'} \mathcal{L}_{A_i^{val}}(f_{\theta_i'}). \quad (7)$$

3 Experiments

3.1 Dataset

We use Assamese, Tamil, and Marathi datasets from Government of India DeitY-TDIL and Bengali, Sinhala and Nepali datasets (Kjartansson et al., 2018) from Open-SLR. The statistics of the dataset are shown in Table-1. The dataset is imbalanced with languages with a large number of training samples.

3.2 Experimental Details

We preprocess the raw audio inputs into a spectrogram before we fetch it into our conformer based model. Our model utilizes a VGG model (Simonyan and Zisserman, 2015), a 6-layer CNN architecture, as the feature extractor. Our speech recognition model consists of sixteen conformer encoders and three transformer decoder layers with eight heads for multi-head attention. The conformer consists of a $dim_{encoder}$ of 512. In total, our model has around 14.9M parameters. For both the MAML and single training models (training model on target language directly), we end the training process after 3M iterations and 1M iteration respectively. During the fine-tuning step for MAML, we run 15 iterations for each sample. We evaluate our model using a beam search with $\eta = 1$, $\gamma = 0.1$, and a beam size of 5. In the single-training setting as well as MAML based training setting, we downsample the speech data to a 16 kHz audio sample rate. The code can be found at here

We train and evaluate the effectiveness of MAML for Indic languages by comparing its performance with the stand-alone conformer model trained on a single language i.e., single-training setting. For each language in MAML taken as target language during experiment, every other languages are used in training. During testing we fine-tune the MAML with target language and then We evaluate the model performance using the character error rate (CER) and run experiments ten times using different test folds. We report the average and standard error of all folds in the 10%-shot, 25%-shot, 50%-shot, 75%-shot and all-shot settings, where q-shot setting means only q% data is used in training from training set.

LANGUAGE: HINDI

ORIGINAL: मुझे इससे कोई फर्क नहीं पड़ता कि रन कहां बने हैं क्योंकि टेस्ट मैचों में रन तो रन होते हैं
SINGLE: मुझे इसे कोई फर्क नहीं पड़त कि रन कहा बने हैं क्योकि टेस्ट मैचो में रन तो रन होते है
MAML: मुझे इसे कोई फर्क नहीं पड़त कि रन कह बने ह क्योकि टेस्ट मैचो मे रन त रन होते ह

ORIGINAL: बजट तैयार करने में अहम भूमिका होती है आइए जानते हैं बजट तैयार करने वाली टीम के बारे में
SINGLE: बजट तैयार करने में अम भूमिका होती है आए जानते ह बजट तैयार करने वाली टीम के बारे मे
MAML: बजट तयार करने में अम भूमिका होती ह आए जानते बजट तैयार करने वाली टीम के बारे म

LANGUAGE: BENGALI

ORIGINAL: এটি ভারতে অনুষ্ঠিত সর্বকালের বৃহত্তম নির্বাচন।
SINGLE: এটি ভারতে অনুষ্ঠিত সর্বকালের বৃত্তম নির্বচন
MAML: এটি ভারতে অনুষ্ঠত সর্বকালে বৃহত্তম নির্বচন

LANGUAGE: TAMIL

ORIGINAL: இந்தியாவில் இதுவரை நடந்த மிகப்பெரிய தேர்தல் இதுவாகும்.
SINGLE: இந்தியாவில் இதுவர நடந்த மிகப்பெரிய தேர்தல் இதுவாகும்.
MAML: இந்தியாவில் இதுவர நடத மிகப்பெரிய தேரதல் இதுவாகும.

Figure 3: Output of some samples in Hindi, Bengali and Tamil Language for MAML and Single language trained model.

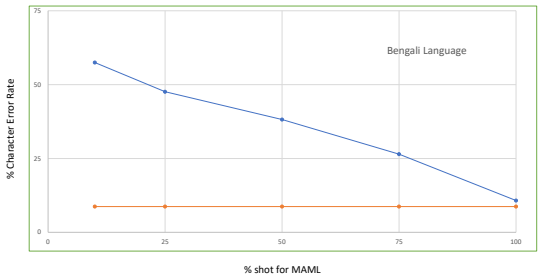

Figure 4: Few-shot results on Bengali Language using MAML vs Single Training.

4 Results and Discussion

4.1 Quantitative Analysis

As shown in Table 2, MAML performance is very close to the model when trained completely on a single language. We have used character error rate (CER) as evaluation criteria because Indic languages contain lot of vowel diacritic which sound similar but are different hence using word error rate (WER) to evaluate will not give correct information on performance of model. Our approach yields up to a 2-4% CER margin in the all-shot MAML and single training. This difference is attributed to low precision in prediction of vowel diacritic for MAML compared to single training. In Figure-3, you can see that sentence generated by MAML is readable and sensible but not entirely correct since there are few missing vowel diacritic.

4.2 Qualitative Analysis

We also evaluate the outputs produced by the model for both MAML based method and the single training method. We evaluate them using a mean human evaluation score that is averaged over 1000 samples for each language. This score is based on the correctness of output and fluency. The scoring is range 0-5 where 0 is for worst performance and 5 for best performance. The evaluation were done by five independent native speakers of each languages. The Table-3 shows the result of the mean human evaluation score for all the languages experimented with.

Few examples generated by both MAML based model and single-language trained model are given in Figure-3.

4.3 Efficacy of Few-Shot Fine-tuning

We investigate the number of samples required to observe performance improvement after fine-tuning the model. We start by training the model with a very small number of samples, i.e., 10%-25% of training data, where each sample approximately consists of 3-4 seconds of audio. We observe that the model cannot adapt to the target language with a such minuscule amount of data. We attribute this to the fact that our model is unable to capture the information from small audio samples due to a large amount acoustic variation in the data. Therefore, we increase the minimum threshold to 10% of the training data, and the model starts to adapt to the target language accordingly. We do this process until the threshold is set to 100%. Figure-4 shows the adaption and constant decrease in CER with an increase in fine-tuning data for the Bengali language.

5 Conclusion

In this paper, we analyse and evaluate the performance of our proposed method for automatic speech recognition in multilingual scenario for seven different low-resource Indic languages. We apply a fast adaptation method on Conformers using model-agnostic meta-learning (MAML) approach to learn a robust automatic speech recognition model to rapidly adapt to unseen languages. Based on the empirical results, MAML consistently comes close to single trained model using target unseen language with a margin of 2-4% CER in all such low-resource multilingual scenarios for Indic languages.

6 Acknowledgement

The work is done in collaboration with NVIDIA. We thank the support from Nvidia, India for providing the computing power and compute infrastructure requirements along with the software stack for the project. We would also like to thank our colleagues at IIIT Guwahati and ISI Kolkata for their valuable input during manual domain language analysis.

7 Note

The work was done when Kumari Deepshikha was at NVIDIA.

References

Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545.

J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiát, S. Watanabe, and T. Hori. 2018. Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 521–527.

S. Dalmia, R. Sanabria, F. Metze, and A. W. Black. 2018. Sequence-based multi-lingual low resource speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4909–4913.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia. PMLR.

Chelsea Finn, Kelvin Xu, and Sergey Levine. 2018. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. 2020. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv preprint arXiv:2005.03191*.

Jui-Yang Hsu, Yuan-Jui Chen, and Hung yi Lee. 2019. Meta learning for end-to-end low-resource speech recognition.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. 2018. Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 52–55, Gurugram, India.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. 2019. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. *arXiv preprint arXiv:1910.10261*.

Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M Cohen, Huyen Nguyen, and Ravi Teja Gadde. 2019. Jasper: An end-to-end convolutional neural acoustic model. *arXiv preprint arXiv:1904.03288*.

Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran. 2013. Deep convolutional neural networks for lvcsr. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8614–8618. IEEE.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Sibo Tong, Philip N. Garner, and Hervé Bourlard. 2017. An investigation of deep neural networks for multilingual speech recognition training and adaptation. In *Proceedings of Interspeech*, pages 714–718, Stockholm, Sweden.

Sibo Tong, Philip N. Garner, and Hervé Bourlard. 2018. Multilingual training and cross-lingual adaptation on ctc-based acoustic model.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard. 2014. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7639–7643.

G. I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, and P. Fung. 2020. Lightweight and efficient end-to-end speech recognition using low-rank transformer. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6144–6148.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280.

Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. 2020. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss.