



SRI-B End-to-End System for Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages

Hardik Sailor, Kiran Praveen T, Vikas Agrawal, Abhinav Jain, Abhishek Pandey

Samsung R&D Institute, Bangalore (SRI-B), India

h.sailor@samsung.com, k.praveen.t@samsung.com, vik.agrawal@samsung.com, jain.abhinav@samsung.com, abhi3.pandey@samsung.com

Abstract

This paper describes SRI-B’s end-to-end Automated Speech Recognition (ASR) system proposed for the subtask-1 on multilingual ASR challenges for Indian languages. Our end-to-end (E2E) ASR model is based on the transformer architecture trained by jointly minimizing Connectionist Temporal Classification (CTC) & Cross-Entropy (CE) losses. A conventional multilingual model which is trained by pooling data from multiple languages helps in terms of generalization, but it comes at the expense of performance degradation compared to their monolingual counterparts. In our experiments, a multilingual model is trained by conditioning the input features using a language-specific embedding vector. These language-specific embedding vectors are obtained by training a language classifier using an attention-based transformer architecture, and then considering its bottleneck features as language identification (LID) embeddings. We further adapt the multilingual system with language specific data to reduce the degradation on specific languages. We propose a novel hypothesis elimination strategy based on LID scores and length-normalized probabilities that optimally select the model from the pool of available models. The experimental results show that the proposed multilingual training and hypothesis elimination strategy gives an average 3.02% of relative word error recognition (WER) improvement for the blind set over the challenge hybrid ASR baseline system. **Index Terms:** Indian languages, language embedding, transformer architecture, multilingual ASR, hypothesis elimination

1. Introduction

Recently, there has been a significant rise in interest for the development of multilingual systems for speech recognition, especially in countries with high linguistic diversity such as India, which has more than 1652 native languages/dialects, 22 of which are officially acknowledged. Owing to this diversity, most of the Indian languages are low resourced, compared to the likes of other languages speech corpora used for developing Automatic Speech Recognition (ASR) systems. To induce advancements and innovations in ASR for Indian languages, the ASR challenge for three low resource Indian languages was organised as a special session during Interspeech 2018 [1]. In continuation, recently another multilingual and code-switching ASR challenge was organised targeting six Indian languages [2].

Recent studies on ASR for Indian languages [1] have shown that the standard approach which trains a single model by pooling all the available data tends to perform well only on a subset of the languages that were used for training, the reason for which could be the confusions caused by the similarities among the languages. Studies have also shown that learning generalised representations for all the languages lead to degradation

in performance [3]. To combat this issue, the model could be provided with language-specific information, which might help the model to differentiate among similar languages. There are multiple ways of incorporating this information into the model such as using a multitask loss where the secondary task is classifying the language, or using language embeddings which are obtained from a standalone language classifier. It has been shown that using an external LID classifier is more beneficial than joint training of LID and ASR systems [4], a possible explanation being the LID loss not contributing much during the process of training the model, as the magnitude of this loss is small compared to that of the cross-entropy or CTC loss. Therefore, this work relies on an external LID classifier to generate language embeddings.

In this paper, we use an end-to-end multilingual ASR model that takes in Mel filterbank features with LID embedding vectors super-imposed at each frame.

Our key contributions in this paper are as follows:

- A standalone LID classifier using the transformer architecture [5] is trained to learn embeddings that are used to condition the ASR input features
- A novel hypothesis elimination technique is proposed using LID scores and length-normalized probabilities to take advantage of multiple models.

2. Related Work

Multilingual ASR is a promising approach towards building an ASR system for low resource languages as information can be shared across languages, which may help in improving the performance. Recently, end-to-end ASR systems have shown better or comparable results compared to that of hybrid DNN-HMM systems for this task. The first study of using an end-to-end model for multilingual ASR is reported in [6] which is based on a hybrid CTC/attention model trained by pooling the speech data in 10 languages, where language information is provided by adding language tags in the transcription of each language. In [7], language embeddings are provided to both the encoder and decoder of a Listen Attend and Spell (LAS) architecture, for developing a single multilingual model. Stacked bottleneck features from hybrid DNN-HMM are used for end-to-end ASR training to effectively combine traditional DNN and Seq-to-Seq model in [8].

In [9], language-specific CTC losses are added with shared hidden layers instead of a single CTC model. The same architecture is utilized in [10] along with a sampling strategy to take advantage of the corpus relatedness. With a similar architecture as [9], approaches in [11], [12], [13] and [14] have also added language-specific gating units in the model. Primary CE loss with multilingual context-dependent phonemes and auxiliary CE loss using language-specific phonemes was

proposed in [15]. There are some approaches that are based on a transliteration scheme to make the model language-agnostic [16], [17]. Byte-level representation was shown to perform better than character-based units in the LAS model in [18]. Language-specific residual adapters were proposed using streaming RNNT models in [19]. Recently, many multilingual models were developed for streaming applications [17, 19, 20]. More recently, mixture of experts model is proposed in [21] to assign per-language parameters in the multilingual model. Several configurations for using LID were explored in [4, 22].

System combination is a well explored area in conventional ASR models since multiple ASR model may have complementary information, combing them using appropriate techniques generally improves the performance. However, it is not explored to a great extent in end-to-end ASR approaches. In [23], a hypothesis-level combination between the hybrid DNN-HMM model and end-to-end DNN models was proposed. MBR training and length normalized scores were used for hypothesis combination. In this work, we combine hypotheses from adapted monolingual and multilingual models using LID scores and length normalised scores.

The overall organization of the paper is as follows: Section 3 contains the details of the database used in this study. Our end-to-end multilingual ASR model is described in Section 4. Our proposed hypothesis elimination approach is described in Section 5. Experimental results are reported in Section 6 and the paper is summarized in Section 7.

3. Data Description

All experiments are performed using the multilingual speech data which is provided by the organizers of multilingual and code-switching ASR challenges [2]. The dataset consists of 6 Indian languages Hindi (HI), Marathi (MA), Odia (OD), Gujarati (GU), Tamil (TA), and Telugu (TE), the details of which are shown in Table 1. Henceforth, we use language names or abbreviations interchangeably. Hindi and Marathi data include utterances from a collection of stories, while Odia has utterances from healthcare, agriculture & finance domains, and Gujarati, Tamil, and Telugu data are from general categories. The development set for the individual languages is also released along with the training data and the blind set is released at the later stage of the challenge that consists of the audios from all 6 languages without any language information. A detailed description of the datasets is mentioned in [2].

Table 1: Details of multilingual database

	GU	TA	TE	HI	MA	OD
Train (hrs)	40	40	40	95.05	93.89	94.54
Dev (hrs)	5	5	5	5.55	5	5.49
# Utts in train	22807	39131	44882	65950	52288	39188
# Utts in dev	3075	3081	3040	3843	4675	3471
# Words	41238	57883	48686	6053	3246	1584

4. End-to-End Multilingual ASR

Our multilingual transcription model uses the hybrid CTC/attention based end-to-end transformer architecture, composed of two main blocks: the encoder and the decoder. A multi-objective learning framework as proposed in [24] is used to jointly train the network with both the losses. After model

training, joint decoding is performed using CTC and attention decoding.

4.1. Conditioning Using Language Embedding

In this model, a language embedding is used as a way of conditioning the model. The input features are super-imposed with the language embeddings, repeated at every frame. For identifying the language of any utterance, we utilise a stand-alone language classifier network, whose architecture is shown in Figure 1. The LID network consists of attention-based encoder layers and a fully connected bottleneck layer with the equal dimensions as input features in the transcription network. The final layer is a fully connected output layer of 6 dimensions (One for each class). After training, the bottleneck features (BNF) obtained from this classifier are used as language embeddings. In this paper, BNF and language embedding are used interchangeably.

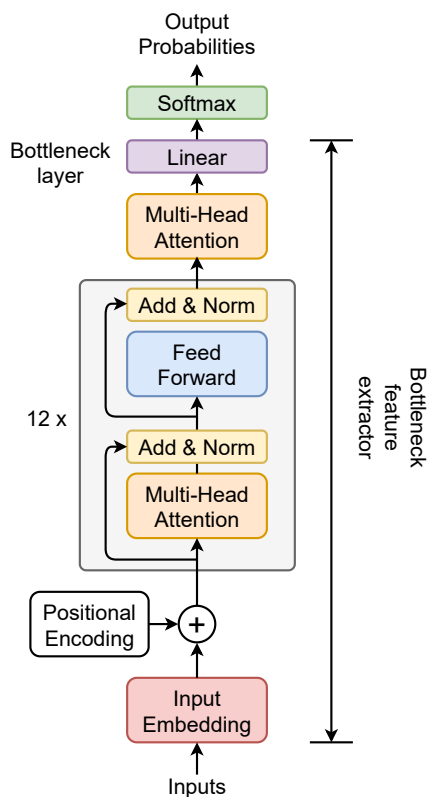


Figure 1: Language classifier network

We incorporate the language information in our transcription network using the BNF from our LID classifier. For this, every utterance is passed through the language classification network and the bottleneck features are repeated and added to every input frame of the corresponding utterance. The architecture of the transcription network with language embeddings is shown in Figure 2.

4.2. Details of the E2E ASR system

All end-to-end ASR models are trained using the ESPnet toolkit [25]. For all the experiments, we use 83-dimensional features that include 80-dimensional Mel filterbank features and 3-dimensional pitch features. The features are extracted at 25ms

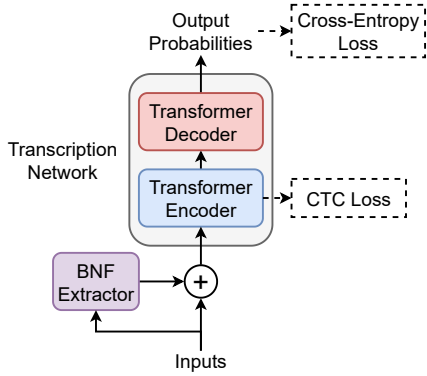


Figure 2: BNF conditioned multi-lingual ASR

with a Hamming window and a hop of 10ms. The model we use for our ASR is based on the transformer architecture [5]. Both the LID and ASR models have 12 encoder layers, while the ASR model also has 6 decoder layers. The LID classifier also uses the same 83-dim features to get attention bottleneck as language embeddings. The output targets are sub-words generated by training a unigram on the text. Monolingual ASR models for GU, TA, and TE are trained using 2k sub-word units while HI, MA, and OD models use 1k sub-words. The multilingual models use 6k sub-word units segmented from pooling the data from all the languages. Our models adopt the standard hybrid CTC/attention training [24] mechanism with a CTC weight of 0.3. For decoding, we use the beam search algorithm with a beam size of 12 with the CTC weight set to 0.4. We do not perform any language model rescoring for our experiments.

5. Hypothesis Elimination

We select the final hypothesis considering a total of 8 models - 6 adapted multilingual models for each language and 2 multilingual (one with language embeddings and one without). The motivation behind choosing these models is as follows,

- Multilingual models provide adequate accuracy on average. However, it was observed from the validation set that the multilingual model without LID is more suited for the languages GU, TA, and TE, and the one with LID is more suited for the remaining languages. We could utilize the language classifier for getting the best out of both models.
- During experiments, we have observed that tokens were getting mixed from the multiple languages for the multilingual model. The multilingual model is further adapted with the individual target language data to resolve this issue. Although, using monolingual models and language classifier alone will increase the WER since the classifier is only 95.38% accurate.
- Exploiting the strengths of all the models could help in determining the final hypothesis.

We consider the normalised log probability of a hypothesis as a weak indicator of its WER, as it was observed that the WER and the normalized log probability of the hypotheses are negatively correlated. Using a validation set we compute Pearson's correlation coefficient between the normalised log-probability and its WER, and find this to be negative (-0.59), which indicates that a higher normalised log-probability generally implies a better

WER. Normalised log probability q is computed as follows:

$$q = \frac{1}{N} \log(P(w_1, w_2, w_3, \dots, w_N)) \quad (1)$$

Where N is the total number of tokens and w_i 's are the optimal output tokens at time i , generated using the beam search. The hypothesis elimination is done separately for every utterance in 2 stages as shown in Figure 3. In the first stage, 5 out of the 6 adapted models are eliminated, along with one of the multilingual models. In the second stage, the final hypothesis is chosen by eliminating one of the chosen models from stage 1.

5.1. Choosing from the adapted language models

For an adapted model of language l , let y_i^* be the best output sequence using beam search. The probability of a language l is denoted by $P(l)$ and is obtained from the language classifier. Let $q_l(y_i^*)$ denote the normalized log-probability of the sequence y in the adapted model with language l . The score S_l for each model l is calculated as follows:

$$S_l = \log(P(l)) + q_l(y_i^*) \quad (2)$$

The model l^* , which proceeds to the next stage is selected as follows:

$$l^* = \arg \max_l S_l \quad (3)$$

5.2. Choosing from the multilingual models

Let M_1 be the multilingual model with LID and M_2 be the one without LID. As it is observed from the validation set that M_2 is more suited for languages GU, TA, and TE, the sum of the probabilities of these languages are assigned to M_2 .

$$P(M_2) = P(l = GU) + P(l = TA) + P(l = TE) \quad (4)$$

$$P(M_1) = 1 - P(M_2) \quad (5)$$

For every utterance, the model with the lower probability is eliminated.

5.3. Choosing the final hypothesis

As mentioned in Section 5.1 and Section 5.2, *Adapted Best* and *Multi Best* respectively are obtained. To eliminate one model for obtaining the final hypothesis, the normalised log-probabilities of *Adapted Best* and *Multi Best* are compared and the model with the lower sequence probability is eliminated.

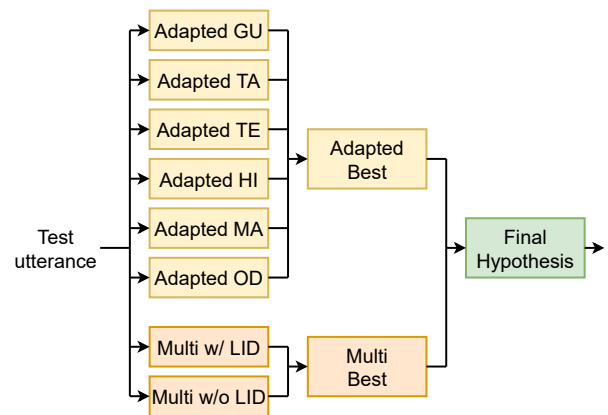


Figure 3: Selecting the final hypothesis

6. Experimental Results

6.1. LID classification

The bottleneck features are generated using the LID classification task. The confusion matrix and accuracies for the classifier shown in Table 2. The languages (HI & MA) is the most confusing pair because these are very similar and even share the same writing scripts. The reason behind confusion among GU, TE & TA could be the similarity in their text category. A similar trend has been seen for the HI, MA & OD.

Table 2: *Confusion matrix and accuracies of the classifier obtained on the development set.*

	GU	TA	TE	HI	MA	OD	Accuracy
GU	1851	1	93	0	0	0	95.17%
TA	6	1898	6	0	0	0	88.34%
TE	48	12	1862	0	0	0	93.84%
HI	0	0	0	2659	346	5	98.70%
MA	0	0	0	215	3307	2	99.37%
OD	3	1	1	14	16	2654	96.88%

6.2. Multilingual ASR results

The experimental results of monolingual and multilingual end-to-end ASR models on development sets are shown in Table 3. The baseline results for this challenge are also shown here for comparison which was trained using a hybrid HMM-TDNN model [2]. The multilingual model significantly reduced WER for HI, MA, and OD. Specifically monolingual OD model has a very high error rate due to overfitting. It should be noted that HI, MA and OD have a very small vocabulary of words compared to the rest of the languages because of which we observe overfitting issues with these three languages. However, the multilingual model mitigates this issue up to a certain extent. Our proposed multilingual model with LID embeddings performs better on HI, MA, and OD with a relative reduction of 4.4 - 10.4 % in WER compared to the multilingual model alone. The primary reason for this improvement is due to conditioning information using LID embeddings. However, this model has degraded the performance of GU, TA, and TE languages. On average, the multilingual model with LID embeddings gives a relative reduction of 2.9 % compared to the multilingual system. This model also has comparable WER with challenge baseline with TDNN hybrid ASR system.

The multilingual model with LID embeddings is used as a pretrained model to adapt to individual languages using their respective monolingual data. The adapted models have the lowest WER for HI and TA languages compared to all the systems. For the rest of the languages, performance lies between the multilingual and multilingual with language embeddings. It can be observed for the results, none of the models give a consistent reduction on WER for all the languages.

Table 3: *Comparison for monolingual and multilingual ASR dev set results along with challenge baseline.*

Model	GU	TA	TE	HI	MA	OD	Avg
Mono	16.8	23.8	37.3	43.4	44.9	82.4	41.4
Multi	18.7	23.7	36.9	37	22.7	51.4	31.7
Multi+LID Emb.	20.0	25.2	38.6	34.9	21.7	44.4	30.8
Adapted	18.9	23.6	37.6	32.0	21.8	46.5	30.1
Challenge Baseline	19.3	33.4	30.6	40.4	22.4	39.1	30.9

6.3. Blind evaluation set results

Experimental results on the blind evaluation set are shown in Table 4. Here, it should be noted that results on the Marathi will be excluded as per the challenge organizers' decision of using a second scoreboard. During blind set submission, it was observed many participants were obtaining significantly high WER on Marathi utterances in the blind set including our results of 76.27 %. Hence, average WER was calculated excluding Marathi set as shown in Table 4. It can be observed that a multilingual model with language embeddings gives an absolute reduction of 6.12, 2.31, and 4.73 % for TA, TE, and HI languages. The model did not perform better for GU and OD. However, the overall performance of 33.49 % is comparable to that result of the challenge baseline.

The results of the hypothesis elimination technique, based on the formulation in Section 5, are also shown in Table 4 referred to as Hyp. Eli. Compared to the multilingual system alone, this technique reduced WER for all the languages except OD. Overall, the hypothesis elimination technique gave an absolute reduction of 1.1 % in WER compared to multilingual with LID embeddings and the challenge baseline system.

Table 4: *Comparison for monolingual and multilingual ASR blind set results without including Marathi set.*

Model	GU	TA	TE	HI	OD	Avg
Challenge Baseline	26.2	34.1	31.4	37.2	38.5	33.5
Multi+LID Emb.	30.2	27.9	29.1	32.5	47.7	33.5
Hyp. Eli.	27.6	26.1	28.3	30.4	49.8	32.4

7. Summary and Conclusions

In this paper, we present the development methodology of a multilingual E2E ASR system. This system was submitted for multilingual and code-switch ASR challenges for Indian Languages. Our multilingual model uses the language embeddings obtained from the transformer-based LID classifier. The proposed system with a language information performs better on an average as compared to the multilingual model without LID. We also propose the novel hypothesis elimination method to exploit the complementary information learned by different models. With the combination of both of the aforementioned techniques, the proposed system performs better than our E2E multilingual system and a baseline hybrid ASR system from the challenge.

8. Acknowledgments

We would like to acknowledge organizers of the challenges on multilingual and code-switch ASR in Indian languages to provide the database. We sincerely thank Samsung R&D Institute, Bangalore for the support to carry out this work.

9. References

- [1] B. M. L. Srivastava, S. Sitaram *et al.*, "Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages," in *Proc. The 6th Intl. Workshop on SLTU*, 2018, pp. 11–14.
- [2] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra, A. Mittal, P. K. Ghosh, P. Jyothi, K. Bali, V. Seshadri, S. Sitaram, S. Bharadwaj, J. Nanavati, R. Nanavati, K. Sankaranarayanan, T. Seeram, and B. Abraham, "Multilingual and code-switching asr challenges for

- low resource Indian languages,” *Proc. Interspeech, Brno, Czech Republic*, 2021.
- [3] A. Kannan *et al.*, “Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model,” in *Interspeech*, 2019, pp. 2130–2134.
- [4] S. Punjabi *et al.*, “Joint ASR and language identification using RNN-T: An efficient approach to dynamic language switching,” in *IEEE ICASSP*, 2021.
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, 2017.
- [6] S. Watanabe, T. Hori, and J. R. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in *ASRU*, Dec 2017, pp. 265–271.
- [7] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, “Multilingual speech recognition with a single end-to-end model,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4904–4908.
- [8] M. Karafiát, M. K. Baskar, S. Watanabe, T. Hori, M. Wiesner, and J. Černocký, “Analysis of Multilingual Sequence-to-Sequence Speech Recognition Systems,” in *Interspeech*, 2019, pp. 2220–2224.
- [9] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, “Sequence-based multi-lingual low resource speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4909–4913.
- [10] X. Li, S. Dalmia, A. W. Black, and F. Metze, “Multilingual Speech Recognition with Corpus Relatedness Sampling,” in *Proc. Interspeech*, 2019, pp. 2120–2124.
- [11] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, “Joint modeling of accents and acoustics for multi-accent speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1–5.
- [12] S. Kim and M. L. Seltzer, “Towards language-universal end-to-end speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4914–4918.
- [13] D. Liu, X. Wan, J. Xu, and P. Zhang, “Multilingual speech recognition training and adaptation with language-specific gate units,” in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 86–90.
- [14] Y.-F. Liao, M. Pleva, D. Hladek, J. Stas, P. Vizlay, M. Lojka, and J. Juhar, “Gated module neural network for multilingual speech recognition,” in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 131–135.
- [15] H. B. Sailor and T. Hain, “Multilingual Speech Recognition Using Language-Specific Phoneme Recognition as Auxiliary Task for Indian Languages,” in *Proc. Interspeech 2020*, 2020, pp. 4756–4760.
- [16] S. Thomas, K. Audhkhasi, and B. Kingsbury, “Transliteration based data augmentation for training multilingual ASR acoustic models in low resource settings,” *Proc. Interspeech 2020*, pp. 4736–4740, 2020.
- [17] A. Datta, B. Ramabhadran, J. Emond, A. Kannan, and B. Roark, “Language-agnostic multilingual modeling,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8239–8243.
- [18] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, “Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5621–5625.
- [19] Y. Zhu, P. Haghani, A. Tripathi, B. Ramabhadran, B. Farris, H. Xu, H. Lu, H. Sak, I. Leal, N. Gaur *et al.*, “Multilingual speech recognition with self-attention structured parameterization,” *Proc. Interspeech 2020*, pp. 4741–4745, 2020.
- [20] A. Waters, N. Gaur, P. Haghani, P. Moreno, and Z. Qu, “Leveraging language ID in multilingual end-to-end speech recognition,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 928–935.
- [21] B. Ramabhadran, B. Farris, I. Leal, M. Prasad, N. Gaur, P. Haghani, P. J. M. Mengibar, and Y. Zhu, “Mixture of experts for multilingual speech recognition,” in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2021.
- [22] V. M. Shetty, M. Sagaya Mary N J, and S. Umesh, “Improving the performance of transformer based low resource speech recognition for Indian languages,” in *IEEE ICASSP*, 2020, pp. 8279–8283.
- [23] J. Wong, Y. Gaur, R. Zhao, L. Lu, E. Sun, J. Li, and Y. Gong, “Combination of end-to-end and hybrid models for speech recognition,” in *Interspeech*. ISCA, October 2020.
- [24] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid CTC/Attention architecture for End-to-End speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
- [25] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “ESPnet: end-to-end speech processing toolkit,” in *Proc. Interspeech*, 2018, pp. 2207–2211.