Interspeech 2024
1-5 September 2024, Kos, Greece

Rapid Language Adaptation for Multilingual E2E Speech Recognition Using Encoder Prompting

Yosuke Kashiwagi[1], *Hayato Futami*[1], *Emiru Tsunoo*[1], *Siddhant Arora*[2], *Shinji Watanabe*[2]

[1]Sony Group Corporation, Japan, [2]Carnegie Mellon University, USA

yosuke.kashiwagi@sony.com

Abstract

End-to-end multilingual speech recognition models handle multiple languages through a single model, often incorporating language identification to automatically detect the language of incoming speech. Since the common scenario is where the language is already known, these models can perform as language-specific by using language information as prompts, which is particularly beneficial for attention-based encoder-decoder architectures. However, the Connectionist Temporal Classification (CTC) approach, which enhances recognition via joint decoding and multi-task training, does not normally incorporate language prompts due to its conditionally independent output tokens. To overcome this, we introduce an encoder prompting technique within the self-conditioned CTC framework, enabling language-specific adaptation of the CTC model in a zero-shot manner. Our method has shown to significantly reduce errors by 28% on average and by 41% on low-resource languages.
Index Terms: speech recognition, E2E, multi-lingual, prompting, adaptation

1. Introduction

In recent years, multilingual speech recognition models have emerged, taking advantage of massive computational resources and large amounts of data [1–8]. These are efficiently trained by aggregating large amounts of data into a single model. For example, Whisper supports more than 100 languages in a single model [3]. Open Whisper-style speech model is a model that reproduces a Whisper-like model with open data [4, 5]. Google universal speech model uses large amounts of unpaired data to improve recognition performance for low-resource languages [6]. Meta has proposed MMS and is attempting to extend it to over 1000 languages [7,8]. Multilingual speech recognition has been reported to be advantageous in terms of data efficiency, especially in low-resource languages [9, 10].

Language identification is often provided as a secondary function of multilingual speech recognition. In addition to simply identify the language of speech, there are also attempts to detect language switches in multi-speaker conversations [11]. However, language identification is not always required for daily use in speech recognition systems. When individuals use speech recognition, the language they speak is often predetermined. E2E multilingual speech recognition provides a function to adapt the model for a specific language by giving the target language ID as a prompt to the decoder, especially for the attention-based encoder decoder [3]. It has been reported that providing a language ID can significantly improve multilingual recognition performance [12–17].

On the other hand, it is also reported that multi-task training and joint-decoding using Connectionist Temporal Classification (CTC) can improve recognition performance in E2E multilanguage speech recognition [4, 18, 19]. In this joint-decoding framework, we hypothesize that the performance can be further improved by providing language IDs while computing the CTC output. However, since the CTC output at each frame is conditionally independent, it is not possible to adapt the recognition results by providing a language ID as the decoder.

Self-conditioned CTC (SC-CTC) is proposed to mitigate the conditional independence of CTC. SC-CTC calculates CTC loss in the middle layer of the encoder and adds the intermediate prediction to the input of the next encoder layer [20]. Many variants of SC-CTC have been proposed to improve its performance [21–23]. Related to multilingual speech recognition, hierarchically changing the importance of language tokens during training has also been proposed [24]. However, this previous study targets training methods and not language adaptation during inference.

Our proposed method uses the general SC-CTC framework during training and can be adapted quickly by simply providing prompts during inference. Prompting is accomplished by modifying the probabilities of language IDs of the token sequence estimated in the intermediate layer. Prior work adapts by adding linguistic information to the input [23]. However, the prior work requires a significant change in structure during training and does not assume the case where no linguistic information is given. On the other hand, our method can operate as a general self-conditioned model if no prompt is given during inference. We confirmed the effectiveness of our proposed method with the Common Voice, VoxForge and FLEURS corpus [25–27]. Experiments showed that our method achieved an average relative error reduction of 28% on Common Voice data. It was also particularly effective for extremely low-resource languages, achieving an error reduction of 41% for languages with less than 5 hours of training data.

2. Related works

2.1. E2E multilingual ASR

Multilingual speech recognition has received a great deal of attention in recent years, and many studies and models have been published [1–8]. By training many languages with a single model, the recognition performance is improved, especially for languages with a small number of data [9, 10]. In our study, we further focus on the E2E multilingual model, which estimates language IDs simultaneously with the text. As shown in Figure 1, this model uses the token sequence with the language ID added to the beginning of the transcription in training. During inference, users can choose to explicitly specify the language or not. If they want to specify the language, they can fix the language ID; if not, they can use the model to infer the language ID

10.21437/Interspeech.2024-702

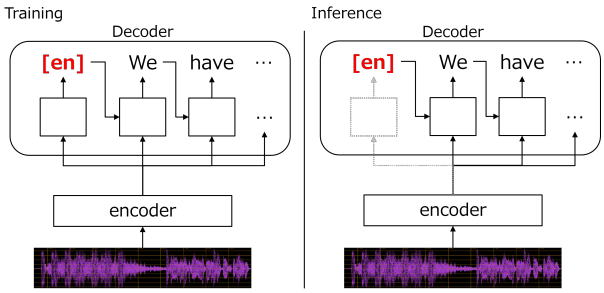

Figure 1: *Multilingual ASR with language prompts added to the decoder.*

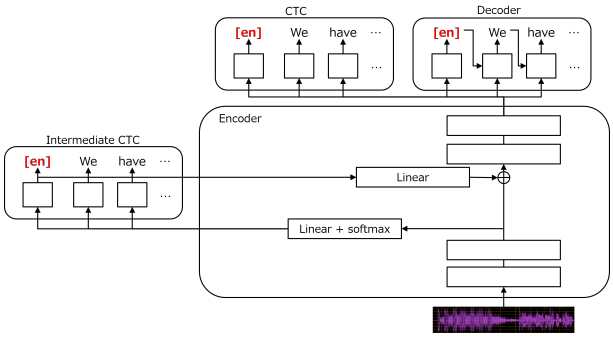

Figure 2: *Joint CTC-AED model with self-conditioned CTC for multilingual ASR.*

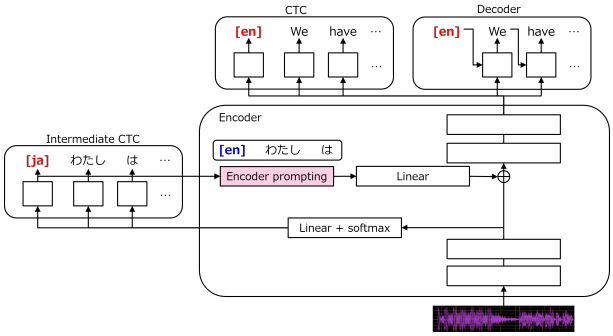

Figure 3: *Proposed rapid language adaptation with encoder prompting during inference.*

as well. This allows the model to be flexibly controlled according to the situation in which it is used. For example, in situations where the input speech comes from an unspecified number of speakers, such as on the internet, supporting as many languages as possible will provide many people with the opportunity to interact freely. On the other hand, in situations where the language of input speech is predetermined, such as on a personal device, the recognition accuracy can be improved by using the system specialized for a particular language.

2.2. Self-conditioned CTC

CTC is a non-autoregressive model in which each frame is conditionally independent [18]. This is an advantage of fast computation, but it can also be a disadvantage. Since conditional independence is a strong assumption, generating text sequences using CTC alone can be suboptimal. Therefore, CTC is often used in combination with an attention decoder [19]. It has been reported that multitask training and joint-decoding with CTC are effective for E2E speech recognition. This technique is useful not only for improving recognition accuracy but also for contributing to training stability. On the other hand, SC-CTC has been proposed to mitigate the conditional independence assumption [20]. SC-CTC is an extension of intermediate CTC (InterCTC) [28]. InterCTC is a method of performing multitask training by calculating CTC losses in the intermediate layer of the encoder. This has been shown to stabilize encoder training. SC-CTC further adds the estimated CTC labels at the intermediate layer to the next layer through a linear transformation. It has been reported that this subnetwork improves the recognition performance of the CTC model. Furthermore, SC-CTC improves the joint-decoding performance using CTC and an attention-based encoder-decoder (AED) for multilingual ASR as shown in Figure 2 [24].

3. Proposed method

3.1. Rapid language adaptation

There are two ways to use E2E multilingual speech recognition. One is to perform language identification and speech recognition simultaneously, which is targeted for environments where speech from an unspecified number of speakers is input. The other is to use the model with a predefined language, which is intended for use on personal devices or within a limited community scenario where the language of the speaker is predetermined. We propose encoder prompting, a new rapid language adaptation method targeting the latter case where the input language is predetermined. The encoder prompting is to control not only the decoder but also the encoder in a prompt-based manner. To achieve this, a SC-CTC-based joint ASR CTC model is first trained, as shown in Figure 2. During inference, language IDs are modified to the predefined target language in the intermediate layer of the SC-CTC, as shown in Figure 3 (Encoder Prompting). Sec. 3.2 describes how to modify the labels. After that, the modified token sequence is returned to the middle layer through a linear transformation similar to SC-CTC. In Figure 3, the input speech is English, but the intermediate layer has the highest probability of being Japanese. In this case, the model knows the input language is English because it is predetermined by the user. Therefore, the output of the intermediate layer is modified so that the probability of an English token is the highest. Since the proposed method requires modification only during inference, it can be used to quickly adapt an already trained model. Furthermore, the modified token sequence affects the output of the encoder, so it is reflected in the output of both the attention decoder as well as the CTC.

3.2. Encoder prompting

We investigated three ways to modify the output of CTC to reflect the language information. The first is to modify only the frames with the highest language ID probability as:

$$\hat{p}_t(k) = \begin{cases} \mathrm{OneHot}(k_{\mathrm{target}}) & \arg\max_{k'} p_t(k') \in K_{\mathrm{LID}} \\ p_t(k) & \text{otherwise} \end{cases}, \quad (1)$$

where $p_t(k)$ is the intermediate layer's output probability for token k in time t and $\hat{p}_t(k)$ is the modified probability after encoder prompting. K_{LID} is the set of tokens corresponding to the language ID and k_{target} is the target language ID. $\mathrm{OneHot}(k_{\mathrm{target}})$

Table 1: *Comparison of related baseline models and our proposed encoder prompting (Sec. 3.2) on Common Voice evaluation set in CER(%). The languages are divided into high, middle, low, and extreme low according to the amount of data. High is for languages with more than 100 hours of data, middle is for 20-100 hours, low is for 5-20 hours, and extremely low is for less than 5 hours.*

ID	encoder type	joint decoding	decoder prompting	encoder prompting	high	middle	low	ex.low	avg.
(a)	Transformer				6.1	10.5	26.7	39.7	22.9
(b)	Transformer	✓			6.2	10.3	26.2	39.6	22.6
(c)	InterCTC	✓			6.1	11.4	28.1	40.6	23.7
(d)	SC-CTC	✓			6.1	11.2	27.9	42.5	24.2
(e)	Transformer		✓		5.9	9.3	21.4	35.5	19.9
(f)	Transformer	✓	✓		6.1	9.7	23.5	35.0	20.4
(g)	InterCTC	✓	✓		6.0	10.5	25.5	35.9	21.3
(h)	SC-CTC	✓	✓		6.1	10.4	26.2	38.7	22.4
(i)	SC-CTC	✓	✓	*Replacement*	**5.8**	**8.7**	**18.1**	**20.9**	**14.3**
(j)	SC-CTC	✓	✓	*Aggregation*	**5.8**	**8.7**	**18.1**	21.0	**14.3**
(k)	SC-CTC	✓	✓	*Prefix*	6.1	10.4	26.2	38.7	22.4

Table 2: *Language ID accuracy (%) for various encoders. The ID of each row are unified with Table 1.*

ID	encoder type	high	middle	low	ex.low
(b)	Transformer	98.2	93.0	78.5	43.6
(c)	InterCTC	97.7	89.9	73.8	38.4
(d)	SC-CTC	97.7	90.4	73.6	33.9

describes the Kronecker delta. If $k = k_{\text{target}}$, it will be 1; otherwise it will be 0. In this approach, the CTC output frames representing the language information are overwritten by the appropriate target language using a one-hot vector. We call this approach as *Replacement*. Although this approach requires minimal modification, it does not modify the linguistic information that would have remained in other frames. This is because the probability of language IDs is maximal only at the beginning of an utterance. However, language information that remains in other frames, even with low probability, may affect recognition performance.

Therefore, we further propose *Aggregation* approach that modifies Eq.(1) to aggregate the probabilities of the other language IDs to the appropriate language ID in all frames as:

$$\hat{p}_t(k) = \begin{cases} p_t(k) & k \notin K_{\text{LID}} \\ \sum_{k' \in K_{\text{LID}}} p_t(k') & k = k_{\text{target}} \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

This approach can correct the language information in all frames. Since the same process is applied to all frames, the implementation is simple. However, *Aggregation* may result in the probability of language IDs being too high in unintended frames. If this were the case, it would be an unnatural label for multilingual speech recognition.

On the other hand, unlike decoder prompts, *Replacement* and *Aggregation* do not allow the input of natural language prompts of any length. To mitigate this, as a third approach, we propose *Prefix* method that overwrites only the minimum number of frames necessary to represent the prompt from the beginning of the utterance. In our experiments, language ID requires only one frame. Therefore, we modify Eq.(1) such that only the head frame is replaced as:

$$\hat{p}_{t=0}(k) = \text{OneHot}(k_{\text{target}}). \quad (3)$$

Table 3: *Investigation of using soft prompting (Sec. 3.3) on Common Voice data in CER(%). In all conditions, encoder is SC-CTC and joint-decoding is used. The soft prompting is applied to 3 languages during inference (English, Chinese and Japanese). The ID of each row are unified with Table 1.*

ID	decoder prompting	encoder prompting	EN	CN	JA
(h)	✓		8.5	23.4	48.2
(j)	✓	*Aggregation*	8.4	22.6	36.5
(d)			8.5	25.0	51.2
(l)		*Soft aggregation*	**8.4**	**23.2**	**43.4**

3.3. Soft prompting

There may be applications where it is not necessary to recognize all languages, but only a few. For example, in a service where three languages (e.g. English, Chinese and Japanese) are assumed to be input, it is necessary to perform language identification among the three languages. The condition that only three languages are input out of more than a hundred language candidates is beneficial for speech recognition. Our encoder prompting can provide this information to the model. We propose to extend the encoder prompting approach, *Aggregation* (Eq.(2)), specifically to multiple languages as:

$$\hat{p}_t(k) = \begin{cases} p_t(k) & k \notin K_{\text{LID}} \\ \frac{\sum_{k' \in K_{\text{LID}}} p_t(k')}{\sum_{k' \in K_{\text{target}}} p_t(k')} p(k) & k \in K_{\text{target}} \\ 0 & \text{otherwise} \end{cases}, \quad (4)$$

where $K_{\text{target}} \subseteq K_{\text{LID}}$ represents the set of target language candidates. If the number of candidates is 1 then this is equivalent to Eq.(2). There is an approach to input language code in multiple languages in prior work, though, but it requires specialized structure for adaptation [29]. They use the one-hot-vectors assigned to each language, normalized to add up to a total of 1. However, it requires the network to be trained to input language codes. Our approach can be realized simply by using the general SC-CTC at training time and adapting it using Eq. 4 during inference.

Table 4: *Comparison on FLEURS evaluation set for each language group in CER(%).*

ID	joint decoding	decoder prompting	encoder prompting	WE	EE	CMN	SSA	SA	SEA	CJK	avg.
(d)	✓			14.2	11.6	12.0	16.4	18.5	18.0	19.5	15.2
(h)	✓	✓		13.9	10.9	11.9	16.2	17.8	17.7	19.5	14.8
(j)	✓	✓	*Aggregation*	**13.7**	**10.4**	**11.7**	**16.1**	**15.8**	**16.8**	**19.4**	**14.2**

4. Experimental evaluation

4.1. Common Voice + VoxForge

We evaluated the effectiveness of the proposed method through experiments on various large scale multilingual datasets. First, we used Common Voice and VoxForge data [25,26]. There were 52 languages in total and 6,000 hours of data in all. In this paper, we categorized the Common Voice evaluation data into four groups of languages based on the amount of data per language. High is for languages with more than 100 hours of data, middle is for 20-100 hours, low is for 5-20 hours, and extremely low is for less than 5 hours. The model sizes used in our experiments were the same except for the intermediate layers. The encoder was a 12-layer transformer, each with 512 dimensions. The number of heads of the attention was 4. The intermediate layer was added at the 6th layer of the encoder, and the weight of the InterCTC was set to 0.3. The number of tokens in the output was 7000, including 52 language tokens. Training was further performed by multi-task training of the CTC and the attention decoder [19]. The CTC weights of the multi-task training were also set to 0.3.

Table 1 shows the results. First, we compared the performance of multilingual models without any prompting ((a) to (d)). Comparing (a) and (b) confirmed that CTC's joint-decoding was also effective in multilingual speech recognition. On the other hand, InterCTC (c) and SC-CTC (d) showed a slight performance degradation. Next, we compared the use of decoder prompting, assuming that the model is being used for a specific language ((e) to (h)). Decoder prompting improved the performance of the multilingual model, and in all cases, performance was improved over the model without decoder prompting. Therefore, as mentioned before, it was confirmed that it was benefical to use decoder prompting appropriately when the input language was known. It was quite interesting to note that when decoder prompting was used as opposed to when not used, the joint decoding degraded the recognition performance. This provides evidence for our hypothesis that the CTC branch was not appropriately controlled by language IDs. Finally, we compared the performance of the proposed encoder prompting ((i) to (k)). *Replacement* (i) and *Aggregation* (j) had almost identical results. Since the output of CTC tends to be sparse, aggregating probabilities and replacing them with one-hot vectors were almost the same operations. On the other hand, *Prefix* (k) had almost no improvement over the model without encoder prompting (h). This was because the original incorrectly estimated language IDs remained in the token sequence by simply modifying the first frame of the utterance. In particular, encoder prompting with *Replacement* and *Aggregation* was able to significantly improve the performance of low-resource languages with relatively poor baseline performance.

Table 2 shows the language identification accuracy. There were no significant differences in performance between the encoder types. However, we observed a tendency for language identification accuracy to decrease with less data. When the language identification performance was low, the impact of encoder prompting was large because the amount of modification by encoder prompting increased.

Table 3 shows the results of the soft prompting. In this experiment, as described in Sec. 3.3, we assumed a situation in which three languages (English, Chinese, and Japanese) were input. In this corpus, English, Chinese, and Japanese are classified as high-, middle-, and low-resource, respectively. The soft prompting was applied to target these three languages. We found that soft prompting reduced errors without explicitly providing language information to decoder prompting.

4.2. FLEURS

We also evaluated on the FLEURS corpus [27]. FLEURS included 102 languages, with a total of 1.4k hours. Therefore, on average, it contained a little over 10 hours of data for each language. In this corpus we also trained a conformer-based SC-CTC. The conformer encoder was a 12-layer, each with 512 dimensions. The number of tokens in the output was 6500, including 102 language tokens. The other settings were same as in Sec. 4.1. Note that the SSL model was not used since we evaluated the correlation of model performance with the amount of data. In this experiments, we used *Aggregation* as encoder prompting, which is simpler to implement than *Replacement*.

Table 2 shows the experimental results. The languages were divided into seven groups as in the previous study [24, 27]. The improvement using the proposed method was smaller for this corpus than for Common Voice. This was partly because the domain of the Common Voice data was open domain, while FLEURS was based on Wikipedia. In this experiment, the average language identification accuracy was 94.2%. This was higher than that of low-resource (5-20 hours) in Table 2, which had a similar amount of data for each language. Therefore, the domain difference made language identification relatively easy. However, we still observed consistent improvements across all language groups, showing the efficacy of our approach.

5. Conclusion

In this paper, we proposed a new multilingual speech recognition adaptation technique using encoder prompting. Encoder prompting allows flexible adaptation during inference in a human-interpretable discrete domain within the encoder. Our approach, built using SC-CTC, improved recognition performance in situations where the language was predefined. Also, unlike decoder prompting, soft prompting can be applied to the encoder even in cases where input can be in any one of multiple languages. We confirmed that recognition performance was also improved in such a setting using soft prompting. There is still a limitation that encoder prompting must maintain the estimated token sequence length, which is CTC-dependent. It may be possible to mitigate this constraint by using the attention decoder to output the token sequence and return it to the intermediate layer. We believe that controlling the encoder as well as the decoder with prompts should receive more attention.

6. References

[1] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 265–271.

[2] 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 *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4904–4908.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[4] Y. Peng, J. Tian, B. Yan, D. Berrebbi, X. Chang, X. Li, J. Shi, S. Arora, W. Chen, R. Sharma *et al.*, "Reproducing Whisper-style training using an open-source toolkit and publicly available data," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[5] Y. Peng, J. Tian, W. Chen, S. Arora, B. Yan, Y. Sudo, M. Shakeel, K. Choi, J. Shi, X. Chang *et al.*, "OWSM v3. 1: Better and faster open whisper-style speech models based on e-branchformer," *arXiv preprint arXiv:2401.16658*, 2024.

[6] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang *et al.*, "Google USM: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[7] A. Tjandra, N. Singhal, D. Zhang, O. Kalinli, A. Mohamed, D. Le, and M. L. Seltzer, "Massively multilingual ASR on 70 languages: Tokenization, architecture, and generalization capabilities," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[8] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

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

[10] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," *arXiv preprint arXiv:1806.05059*, 2018.

[11] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "End-to-end multilingual multi-speaker speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019, 2019, pp. 3755–3759.

[12] 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)*, 2019, pp. 928–935.

[13] B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, and K. Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4749–4753.

[14] Y. Yang, Y. Peng, X. Zhong, H. Huang, and E. S. Chng, "Adapting OpenAI's Whisper for speech recognition on code-switch mandarin-english seame and asru2019 datasets," *arXiv preprint arXiv:2311.17382*, 2023.

[15] K. Li, J. Li, G. Ye, R. Zhao, and Y. Gong, "Towards code-switching ASR for end-to-end CTC models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6076–6080.

[16] G. I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, P. Xu, and P. Fung, "Meta-transfer learning for code-switched speech recognition," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3770–3776.

[17] S. Dalmia, Y. Liu, S. Ronanki, and K. Kirchhoff, "Transformer-transducers for code-switched speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5859–5863.

[18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[19] T. Hori, S. Watanabe, and J. R. Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 518–529.

[20] J. Nozaki and T. Komatsu, "Relaxing the conditional independence assumption of CTC-based ASR by conditioning on intermediate predictions," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2021. ISCA, 2021, pp. 3735–3739.

[21] T. Komatsu, Y. Fujita, J. Lee, L. Lee, S. Watanabe, and Y. Kida, "Better intermediates improve CTC inference," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022, 2022, pp. 4965–4969.

[22] Y. Nakagome, T. Komatsu, Y. Fujita, S. Ichimura, and Y. Kida, "InterAug: Augmenting noisy intermediate predictions for CTC-based ASR," *arXiv preprint arXiv:2204.00174*, 2022.

[23] S. Li, Y. You, X. Wang, K. Ding, and G. Wan, "Enhancing multilingual speech recognition through language prompt tuning and frame-level language adapter," *arXiv preprint arXiv:2309.09443*, 2023.

[24] W. Chen, B. Yan, J. Shi, Y. Peng, S. Maiti, and S. Watanabe, "Improving massively multilingual ASR with auxiliary CTC objectives," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[26] "Free speech recognition: voxforge.org," https://www.voxforge.org/.

[27] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 798–805.

[28] J. Lee and S. Watanabe, "Intermediate loss regularization for CTC-based speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6224–6228.

[29] L. Zhou, J. Li, E. Sun, and S. Liu, "A configurable multilingual model is all you need to recognize all languages," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6422–6426.