Improving Language Identification for Code-Switched Speech: The Pivotal Role of Accented English

Adyasha Patra, Dhiraj Kumar Sah, Preethi Jyothi
IIT Bombay, India
{adyasha, dhiraj, pjyothi}@cse.iitb.ac.in

Abstract

Code-switching, where speakers alternate between languages within a single utterance, poses unique challenges for language identification (LID). Existing LID models often fail to reliably identify English spoken with the accent of the matrix (dominant) language. We show that finetuning LID models with small amounts of such accented English significantly improves code-switched LID, without degrading performance on standard monolingual speech—a limitation observed with direct finetuning on code-switched utterances. This is achieved via low-rank adaptation (LoRA) on limited accented data, which allows models to adapt efficiently. To better evaluate performance, we introduce LangRank, a metric that captures the relative ranking of identified languages often overlooked by traditional metrics. Our method generalizes across multiple language pairs, including Hindi-English, Bengali-English, Mandarin-English, and Arabic-English, providing robust LID in code-switched multilingual contexts.

1 Introduction

Code-switching, where speakers alternate between two or more languages within a single utterance, is ubiquitous in multilingual communities (Gardner-Chloros, 2009; Nilep, 2006; Myers-Scotton, 2017; Winata et al., 2022). While computational models for code-switching have been developed for tasks such as translation and speech recognition (Winata et al., 2023), spoken language identification (LID) has largely focused on monolingual utterances (Thukroo et al., 2022; O'Shaughnessy, 2024). Predicting utterance-level LID labels for code-switched speech is critical, as it enables routing inputs to language-specific expert models — a mechanism previously explored for improving code-switched ASR (Huang et al., 2024; Wang et al., 2023).

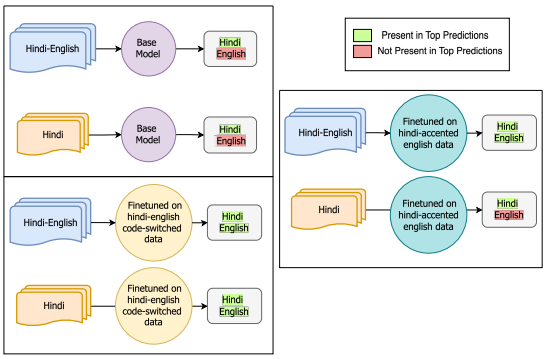

Figure 1: Language identification (LID) models are evaluated on a code-switched Hindi–English and a monolingual Hindi utterance. (a) Without any finetuning, the model does not predict English in a code-switched input. (b) Finetuning on code-switched data causes English to be mistakenly predicted, even for monolingual Hindi. (c) Finetuning on accented English (our proposal) identifies English only when it truly occurs.

How can we adapt pretrained LID models that work well with monolingual speech to perform well on code-switched speech containing English? Ideally, we want the LID model to predict the matrix (dominant) language with highest probability, followed by English. We find that existing LID models struggle with identifying English in code-switched speech, with its detection accuracy being close to zero for Hindi-English and Bengali-English speech.

Our key observation is that existing LID models are confounded by *accented English embedded within the code-switched utterances*. Finetuning pretrained LID models on very small amounts of matrix language-accented English speech improves detection of English, while retaining their base capabilities on the monolingual matrix languages. This approach mitigates overfitting risks associated with finetuning on code-switched data, specifically reducing the likelihood of incorrectly predicting English in purely monolingual matrix-language utterances. Figure 1 illustrates these varying effects

Findings of the Association for Computational Linguistics: EACL 2026, pages 4643–4656

of finetuning on pretrained monolingual LID models.

While accented English still requires targeted collection, it is better represented in public corpora compared to code-switched utterances. For instance, the Mozilla Common Voice speech corpus (Ardila et al., 2020) includes speech representing 16 predetermined English accents and over 200 self-described accents. The Speech Accent Archive (Weinberger, 2015) contains English speech from speakers of 341 different native languages, and the EdAcc corpus (Sanabria et al., 2023) contains more than 40 self-reported English accents. This wider availability of accented English makes it a more scalable finetuning target than code-switched corpora for code-switched LID.

For a more nuanced evaluation of code-switched LID, we also introduce a new metric *LangRank* that captures the relative ranking of predicted languages beyond standard accuracy measures. Our LID approach generalizes across diverse code-switched pairs, including Hindi–English, Bengali–English, Mandarin–English, and Arabic–English.

2 Related Work

Language identification (LID) in speech has been extensively studied, from early acoustic and prosodic feature-based methods (Tong et al., 2006; Zissman and Berkling, 2001) to modern neural architectures (Lopez-Moreno et al., 2014, 2016). However, most prior work has focused on monolingual speech, while research on LID for code-switched speech has primarily targeted detecting switch points or transition boundaries (Nie et al., 2022; Solorio and Liu, 2008; Piergallini et al., 2016; Li et al., 2023).

Large multilingual LID models such as VoxLingua107 (Ravanelli et al., 2021; Valk and Alumäe, 2021), Whisper (Radford et al., 2022), and MMS-LID (Pratap et al., 2023) have advanced language coverage and robustness. Nevertheless, these models still exhibit performance disparities between high-resource and low-resource languages. In addition, Whisper has been observed to overpredict English, sometimes identifying it in purely monolingual non-English utterances.

Recent approaches to code-switched ASR have leveraged Mixture-of-Experts (MoE) architectures to enhance multilingual performance (Huang et al., 2024; Wang et al., 2023). While such methods incorporate language routing, they do not explicitly improve utterance-level language identification, which is critical for reliably directing mixed-language inputs to expert models. Instead, routing typically relies on baseline model logits without targeted adaptation. In contrast, our work directly addresses this limitation by adapting pretrained LID models through finetuning on accented English, enabling more robust & accurate code-switched LID.

3 Methodology

Our task is to predict a set of languages present in a code-switched speech sample at the utterance level, ranked based on the prediction probabilities. This utterance or sentence-level approach aligns with the LID methodologies proposed by Burchell et al. (2024) for text.

We build on the Massively Multilingual Speech (MMS) LID model (Pratap et al., 2023), chosen for its broad language coverage and scalability. MMS-LID is based on the wav2vec 2.0 architecture (Baevski et al., 2020), consisting of 48 Transformer encoder layers, followed by a projection and classification layer. The model maps raw audio to a probability distribution over 126 output classes, where each class represents a language.

To adapt MMS-LID to code-switched speech, we finetune it on limited amounts of accented English audio, ranging from 50 to 500 samples per accent. Given the small size of the finetuning dataset, we employ two parameter-efficient adaptation techniques:

(a) **Adapters** (Chen et al., 2024): Freeze the base MMS-LID layers and insert small adapter layers to efficiently learn new representations without updating the full model.

(b) **Low-Rank Adaptation (LoRA)** (Hu et al., 2021): Apply low-rank updates to the query, key, and value projections in self-attention, enabling adaptation with fewer trainable parameters. Details of these methods & parameters are in Appendix A.

Our experimental design relies on two assumptions: (1) code-switching occurs with English as the embedded language, and (2) the matrix language is known during finetuning to facilitate the selection of the corresponding English accent. Regarding the first assumption, we focus on English-centric code-switching due to the bias in available data; a recent review (Agro et al., 2025) of studies between 2018–2024 shows that Mandarin, Hindi, and Arabic code-switched with English account for ∼76% of all works on computational code-

switching. We think the second assumption of apriori knowledge of the matrix language is reasonable since pretrained LID models are already adept at recognizing the matrix language. Also, choosing an English accent related to the matrix language allows for lightweight finetuning using only 80 samples, which would be insufficient for multi-accent finetuning (as shown in Appendix H).

4 Experimental Setup

4.1 Datasets

For evaluations on code-switched speech, we use Hindi-English & Bengali-English test sets from the MUCS dataset (Diwan et al., 2021), Arabic-English data from the ZAEBUC-Spoken corpus (Hamed et al., 2024), and Mandarin Chinese-English test data from the ASCEND dataset (Lovenia et al., 2022). To check for forgetting in the pretrained LID model, we also create monolingual matrix-language evaluation sets. From the Mozilla Common Voice v13 (Ardila et al., 2020) corpus, we randomly sampled 2000 examples from the test set of each monolingual dataset of Hindi, Bengali, Arabic and Mandarin Chinese. For accented English data, we used Hindi-accented English from NISP dataset (Kalluri et al., 2020), Bengali-accented English from the Svarah dataset (Javed et al., 2023), and Mandarin Chinese & Arabic-accented English from the L2-Arctic dataset (Zhao et al., 2018).

4.2 Evaluation Metrics

4.2.1 Exact Match (EM)

The Exact Match (EM) score measures whether two sets of true and predicted labels exactly match or not. This is a fairly strict metric that was previously used for code-switched LID for text in Kargaran et al. (2024).

In Table 2, we report EM as raw counts of correctly predicted instances (rather than a normalized score) to emphasize performance differences in absolute terms. If there are N instances, and y_i, $\hat{y}_i$ refer to the true and predicted labels for instance i, respectively, then Exact Match (EM) is defined as: $\text{EM} = \sum_{i=1}^{N} \mathbb{I}(y_i = \hat{y}_i)$ where $\mathbb{I}$ is an indicator function that checks whether the sets y_i and $\hat{y}_i$ are an exact match.

4.2.2 LangRank (LR)

Conventional metrics such as *Exact Match (EM)*, *Precision*, and *Recall* (Appendix E) evaluate only whether the ground-truth languages appear among the top k predictions, without considering their relative ordering. For our EM computations, we set $k = 1$ for monolingual evaluation sets and $k = 2$ for code-switched datasets, as the latter involve two target languages. In code-switched speech, this can obscure partially correct predictions: EM requires all ground-truth languages to appear in the top k to count as correct, so languages ranked slightly lower are treated as incorrect even if the model is close to the correct distribution.

For monolingual utterances, EM only considers the single ground-truth language. Over-predictions of other languages (e.g., high probability scores for English) do not affect the score as long as the matrix language is ranked first, failing to capture spurious predictions.

To address these limitations, we propose *LangRank (LR)*, a ranking-based evaluation metric. Let N denote the total number of test sentences, and let $r_{i\ell}$ be the rank of language ℓ in the predicted probabilities for sentence i (rank 1 is highest). LangRank for language ℓ is defined as the reciprocal of its average rank over all test sentences:

$$\text{LR}_\ell = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{r_{i\ell}}.$$

This formulation naturally captures partially correct predictions: languages ranked near the top have higher LR values, while lower-ranked languages have lower LR values. Unlike EM, LR reflects the *relative ordering* of predictions and provides a nuanced evaluation for both code-switched and monolingual utterances (see Table 1 for an illustration).

Let LR_x denote the LangRank of the matrix language and LR_{en} denote the LangRank of English (the embedded language). LR assigns partial credit to each ground-truth language based on its rank in the predicted probability distribution. Higher ranks for relevant languages increase the score, reflecting correct predictions, while lower ranks reduce it. For monolingual utterances, if spurious languages such as English are ranked highly alongside the matrix language, LR_{en} can increase, providing a signal of overfitting.

All implementation details including finetuning specifics, model hyperparameters, etc. are detailed in Appendix B.

Utterance Type	EM	Lang	Rank	LR
Code-switched(Hi-En)	0	Hindi (GT)	1	$LR_x = 1/1 = 1.0$
		English (GT)	3	$LR_{en} = 1/3 \approx 0.33$
		Other	2	-
Monolingual(Hi)	1	Hindi (GT)	1	$LR_x = 1/1 = 1.0$
		English	2	$LR_{en} = 1/2 = 0.5$
		Other	3	-

Table 1: Illustration of EM vs LR for single utterance. See Appendix C for detailed LR calculation of a small toy dataset containing multiple utterances.

5 Experimental Results and Discussion

Oracle Definition. We define an *Oracle* as the ideal prediction reference. For Exact Match (EM), it equals the total number of utterances, i.e., the maximum possible correct predictions. For LangRank (LR), Oracle values depend on the dataset: monolingual sets assume LR_x=1, LR_{en}=0; for code-switched sets, LR is derived using word counts for all languages except Mandarin-Chinese, for which we use character counts (Appendix D).

Exact Match (EM) Performance. Table 2 reports EM accuracy for code-switched utterances across four language pairs: Hindi-English (hi-en), Bengali-English (bn-en), Arabic-English (ar-en), and Mandarin-English (zh-en). Adapter finetuning and the Whisper baseline achieve the highest EM scores across most languages, suggesting strong performance on code-switched data.

System	EM (hi-en)	EM (bn-en)	EM (ar-en)	EM (zh-en)
MMS (Baseline)	509	60	1511	448
MMS (LoRA)	915	401	1617	540
MMS (Adapters)	**2120**	**1407**	1805	789
Whisper (Baseline)	997	893	**2111**	**1054**
Oracle	**3136**	**4275**	**6033**	**1315**

Table 2: EM accuracy for code-switched utterances across four language pairs. Adapters and Whisper yield the highest EM, while the Oracle corresponds to the total number of samples in each dataset.

At this stage, EM suggests that Adapter and Whisper models outperform LoRA. However, as we demonstrate next using LangRank, these systems exhibit severe English overfitting, particularly on monolingual data, revealing that high EM does not necessarily indicate balanced or reliable LID behavior.

LangRank (LR) Analysis. To better capture ranking-based correctness and spurious predictions, we evaluate models using the per-language metric LangRank (LR), introduced in Section 4.2.2. Figure 2 visualizes the trade-off between English LangRank (LR_{en}) on code-switched and monolingual speech. Each point corresponds to a model configuration evaluated across four language pairs.

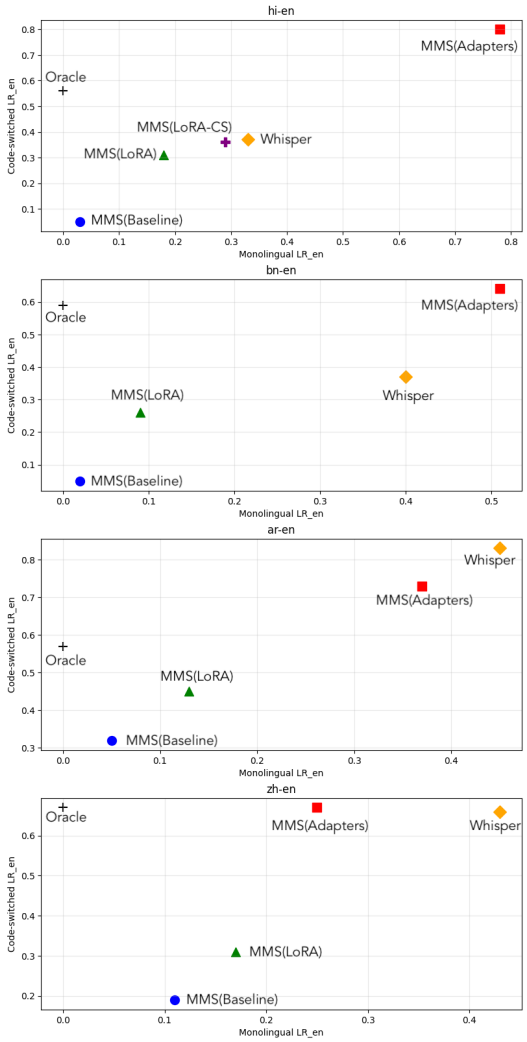

Figure 2: Trade-off between English LangRank (LR_{en}) on code-switched vs monolingual non-English speech. Each subplot corresponds to a different matrix language (Hi, Bn, Ar, Zh). LoRA (green) consistently achieves the smallest Euclidean distance to the oracle ($d_{avg} = 0.31$). Detailed LR values and per-language distances appear in Table 11 (Appendix N).

Adapter-based finetuning and Whisper, despite high EM, over-predict English even in monolingual utterances, producing inflated LR_{en} and deviating sharply from the oracle. By contrast, LoRA finetuning on accented English achieves the closest proximity to the oracle, exhibiting strong English detection when appropriate (in code-switched speech) and restraint in monolingual conditions. Thus, LoRA finetuning on MMS-LID model has the best overall performance.

Failure Analysis and Oracle Gap. Despite the

improvements achieved via LoRA, a performance gap relative to the Oracle persists, as seen in Figure 2 and Tables 8 and 11. Our analysis indicates two primary drivers for these misclassifications. First, acoustic similarity among closely related languages often leads to *distractor* languages ranking higher than English. For instance, in Hindi-English utterances, the model frequently assigns high probability to Urdu or Punjabi due to their phonetic similarities and the speaker's matrix-language accent. This is evidenced by the high LangRank of unrelated but phonetically similar languages in Table 13.

Second, the density of code-switching plays a pivotal role; in low-density utterances, characterized by a low code-mixed index (CMI), a measure of the proportion of non-matrix words in an utterance, the temporal dominance of the matrix language can suppress the English signal, making it difficult for the model to rank English in top-2 even when correctly detected. We leave the integration of switch-point constraints, which could help anchor identification to specific segments, as an area for future work to further narrow this gap.

Impact of Finetuning on Code-switched Data. We finetuned MMS-LID using 80 examples from the Hindi-English code-switched dataset (Rao et al., 2018), following the same LoRA setup as the Hindi-accented English finetuned model. The loss function was modified for this multi-label setting (Appendix F).

In Figure 2, for the Hindi subplot, the model finetuned on Hindi-accented English (denoted as MMS(LoRA)) is closer to the oracle, while the model finetuned on code-switched Hindi-English speech (denoted as MMS(LoRA-CS)) exhibits higher LR_{en} on monolingual Hindi, indicating overfitting. This demonstrates that the observed improvements arise not only from the PEFT technique but also from careful data selection: using accented English for finetuning better balances English detection across code-switched and monolingual contexts.

6 Conclusion and Future Work

For code-switched language identification, we presented a lightweight strategy of adapting pretrained LID models using LoRA finetuning on small sets of matrix-language–accented English samples. This approach significantly improves English detection in code-switched utterances while maintaining its LID performance on monolingual speech. We also introduced *LangRank*, a ranking-based evaluation metric for code-switched LID that reveals tradeoffs obscured by standard accuracy metrics. Our results show that LoRA achieves the best balance between code-switched and monolingual performance, demonstrating its robustness as a targeted adaptation.

Building on these findings, we identify several avenues for exploration. Future work could explore multi-accent extensions like finetuning on a mixture of diverse accented data or accent-invariant representations to improve generalization across varied linguistic backgrounds without requiring prior knowledge of the matrix language. Additionally, future work can aim to extend this framework to non-English code-switched pairs, investigating higher-rank adaptation or contrastive loss formulations to handle cases with significant phonological overlap. We leave the exploration of these non-English contexts to future work as more diverse code-switched corpora become available. Finally, while our preliminary experiments leverage the code-mixed index (CMI) to differentiate performance across mixing extremes, future research should explore this relationship in more detail by evaluating model robustness across a continuous spectrum of mixing densities.

7 Limitations

While this study offers valuable insights into code-switched language identification, we highlight the following limitations. Our approach primarily assumes prior knowledge of the matrix language to select appropriate accented data for adaptation. In real-world contexts where the matrix language may be unknown or misidentified, the effectiveness of this targeted strategy may diminish. Furthermore, while the LoRA-based framework is highly efficient, its benefits are currently accent-specific; our results confirm that performance is sensitive to accent-matched data and does not naturally generalize to mismatched accents.

Finally, the model was evaluated on a limited set of datasets, focusing on English as the embedded language, that does not capture the diversity of code-switching across linguistic communities.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments and construc-

tive feedback, which helped improve the quality of this manuscript. We are also grateful to Prof. Preeti Rao from IIT Bombay for providing the Hindi-English code-switched dataset (Rao et al., 2018) used in our comparative analysis of finetuning data. The last author gratefully acknowledges support from the consortium project on "Speech Technologies in Indian Languages" under National Translation Language Mission (NLTM), MeitY, Government of India.

References

Maha Tufail Agro, Atharva Kulkarni, Karima Kadaoui, Zeerak Talat, and Hanan Aldarmaki. 2025. Code-switching in end-to-end automatic speech recognition: A systematic literature review. *Preprint*, arXiv:2507.07741.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Laurie Burchell, Alexandra Birch, Robert Thompson, and Kenneth Heafield. 2024. Code-switched language identification is harder than you think. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 646–658, St. Julian's, Malta. Association for Computational Linguistics.

Keyu Chen, Yuan Pang, and Zi Yang. 2024. Parameter-efficient fine-tuning with adapters. *Preprint*, arXiv:2405.05493.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *Preprint*, arXiv:2205.12446.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, and Karthik Sankaranarayanan. 2021. Mucs 2021: Multilingual and code-switching asr challenges for low resource indian languages. In *Interspeech 2021*, interspeech_2021. ISCA.

Hugging Face. 2024. Trainer class. Accessed: 2024-10-16.

Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge University Press.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

Injy Hamed, Fadhl Eryani, David Palfreyman, and Nizar Habash. 2024. Zaebuc-spoken: A multilingual multidialectal arabic-english speech corpus. *Preprint*, arXiv:2403.18182.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Hukai Huang, Shenghui Lu, Yahui Shan, He Qu, Fengrun Zhang, Wenhao Guan, Qingyang Hong, and Lin Li. 2024. Dynamic language group-based moe: Enhancing code-switching speech recognition with hierarchical routing. *Preprint*, arXiv:2407.18581.

Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M. Khapra. 2023. Svarah: Evaluating english asr systems on indian accents. *Preprint*, arXiv:2305.15760.

Tahir Javed, Janki Atul Nawale, Eldho Ittan George, Sakshi Joshi, Kaushal Santosh Bhogale, Deovrat Mehendale, Ishvinder Virender Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Sharad Gandhi, Ambujavalli R, Manickam K M, C Venkata Vaijayanthi, Krishnan Srinivasa Raghavan Karunganni, Pratyush Kumar, and Mitesh M Khapra. 2024. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. *Preprint*, arXiv:2403.01926.

Shareef Babu Kalluri, Deepu Vijayasenan, Sriram Ganapathy, Ragesh Rajan M, and Prashant Krishnan. 2020. Nisp: A multi-lingual multi-accent dataset for speaker profiling. *Preprint*, arXiv:2007.06021.

Amir Hossein Kargaran, François Yvon, and Hinrich Schuetze. 2024. MaskLID: Code-switching language identification through iterative masking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 459–469, Bangkok, Thailand. Association for Computational Linguistics.

Shuyue Stella Li, Cihan Xiao, Tianjian Li, and Bismarck Odoom. 2023. Simple yet effective code-switching language identification with multitask pre-training and transfer learning. *Preprint*, arXiv:2305.19759.

Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, David Martinez, Oldřich Plchot, Joaquin Gonzalez-Rodriguez, and Pedro J. Moreno. 2016. On the use of

deep feedforward neural networks for automatic language identification. *Computer Speech & Language*, 40:46–59.

Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno. 2014. Automatic language identification using deep neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5337–5341. IEEE.

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, Xu Yan, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J Barezi, et al. 2022. Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Carol Myers-Scotton. 2017. *Code-Switching*, chapter 13. John Wiley & Sons, Ltd.

Yuting Nie, WeiQiang Zhang, Zhe Ji, and GuiXin Shi. 2022. Language code-switching detection based on bert-lid. In *2022 16th IEEE International Conference on Signal Processing (ICSP)*, volume 1, pages 36–40.

Chad Nilep. 2006. "code switching" in sociocultural linguistics. *Colorado Research in Linguistics*, 19.

Douglas O'Shaughnessy. 2024. Spoken language identification: An overview of past and present research trends. *Speech Communication*, page 103167.

Mario Piergallini, Rouzbeh Shirvani, Gauri Shankar Gautam, and Mohamed Chouikha. 2016. Word-level language identification and predicting codeswitching points in swahili-english language data. In *Proceedings of the second workshop on computational approaches to code switching*, pages 21–29.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

P. Rao, M. Pandya, K. Sabu, K. Kumar, and N. Bondale. 2018. A study of lexical and prosodic cues to segmentation in a hindi-english code-switched discourse. In *Proceedings of Interspeech*, Hyderabad, India.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The edinburgh international accents of english corpus: Towards the democratization of english asr. *Preprint*, arXiv:2303.18110.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981.

Irshad Ahmad Thukroo, Rumaan Bashir, and Kaiser J. Giri. 2022. A review into deep learning techniques for spoken language identification. *Multimedia Tools and Applications*, 81(22):32593–32624.

Rong Tong, Bin Ma, Donglai Zhu, Haizhou Li, and Eng Siong Chng. 2006. Integrating acoustic, prosodic and phonotactic features for spoken language identification. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a dataset for spoken language recognition. In *Proc. IEEE SLT Workshop*.

Wenxuan Wang, Guodong Ma, Yuke Li, and Binbin Du. 2023. Language-routing mixture of experts for multilingual and code-switching speech recognition. *Preprint*, arXiv:2307.05956.

Steven H. Weinberger. 2015. Speech accent archive. `https://accent.gmu.edu/`. Accessed: 2026-01-22.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis,

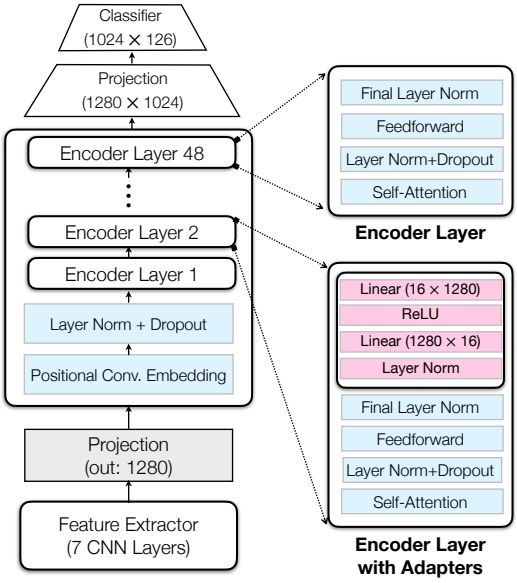

Figure 3: Block structure of the MMS-LID model.

and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech*, page 2783–2787.

Marc A Zissman and Kay M Berkling. 2001. Automatic language identification. *speech communication*, 35(1-2):115–124.

A LoRA and Adapter Details

A.1 Adapter-based Finetuning

In this approach, adapter layers (feedforward modules) are incorporated after every encoder layer to adapt the pretrained MMS-LID model for the new English LID task. By introducing task-specific adapters, we effectively limit the number of trainable parameters, resulting in a model with approximately 3.59 million parameters. These parameters are optimized using the cross-entropy loss function to predict English for a small set of accented English speech samples. The overall architecture of the adapter-enhanced model is illustrated in Figure 3.

A.2 Low-Rank Adaptation (LoRA) Finetuning

The LoRA approach finetunes only the Query (Q), Key (K), and Value (V) projections within each encoder layer using a rank-4 LoRA projection. This method further reduces the number of parameters, compared to adapters, while maintaining the model's ability to learn task-specific behaviour. A detailed representation of the encoder layer and where LoRA adaptation fits in is shown in Figure 4.

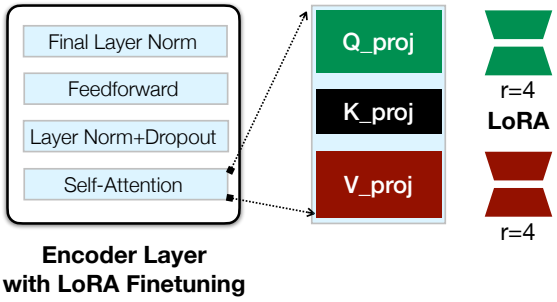

Figure 4: Block structure of an encoder layer in the MMS-LID model, adapted using LoRA finetuning.

The LoRA-based model comprises approximately 1.6 million learnable parameters, which are finetuned using the cross-entropy loss function to predict English for accented English speech samples.

B Implementation Deatils

We use the MMS-LID-126 model as our baseline (Pratap et al., 2023). During data preprocessing, all audio files are processed using the `Wav2Vec2FeatureExtractor` (Baevski et al., 2020), which normalizes the waveform and resamples audio to a sampling rate of 16 kHz.

Adapter Finetuning. Adapter-based finetuning is performed using HuggingFace's `Trainer` API (Face, 2024), with support for mixed-precision training and gradient accumulation.

Low-Rank Adaptation (LoRA). For LoRA-based finetuning, we use a rank of 4 with an alpha scaling factor of 16. All parameter-efficient updates are implemented using the PEFT library (Mangrulkar et al., 2022).

Training Setup. All experiments follow the same training configuration. We use a learning rate of 3×10^{-5}, train for 10 epochs, and apply a warmup of 20 steps. Gradient accumulation is set to 1, training is performed in fp16 precision, and batch size is automatically inferred by the framework, resolving to a batch size of 4 on an NVIDIA RTX A5000 GPU with 24 GB of memory. No early stopping is used.

Finetuning each model on 80 utterances takes approximately 4 minutes on a single RTX A5000 GPU. Unless otherwise stated, all experiments reported in Section 5 use 80 training examples. An ablation study on the effect of finetuning dataset size is presented in Appendix K.

Data Splits. Train and validation splits are fixed across all experiments and explicitly provided in the codebase. For each accent, we use 80 utterances for training and 100 utterances for validation.

Stability and Statistical Reporting. All results are reported as point estimates. Given the low-resource nature of the task, we utilize fixed 80-utterance training sets. Prior studies on parameter-efficient finetuning (PEFT) in low-resource regimes suggest that methods such as LoRA and Adapters exhibit high stability across different initialization seeds (Hu et al., 2021).

To ensure the robustness of our findings, we conducted preliminary experiments with multiple randomly sampled 80-utterance training subsets. We observed minimal variation in the primary metrics (e.g., LR_{en}), with effect sizes consistently exceeding typical run-to-run noise. Consequently, we report results from a representative run for each configuration.

C LangRank(LR) Computation Example

To illustrate the computation of *LangRank (LR)*, consider a test set containing $S = 3$ code-switched utterances. Assume we are evaluating the model's performance for two languages: Hindi (hi) and English (en). For each sentence i, the rank $r_{i,\ell}$ of language ℓ (Hindi or English) is based on the probabilities in the predicted distribution. The ranks are shown in Table 3.

Sentence (i)	**Hindi** ($r_{i,hi}$)	**English** ($r_{i,en}$)
1	1	2
2	3	1
3	2	1

Table 3: Predicted ranks of Hindi and English for three example sentences, used for LR computation.

Using the formula for LR:

$$\mathrm{LR}_\ell = \frac{1}{S} \sum_{i=1}^{S} \frac{1}{r_{i\ell}},$$

we compute LR for Hindi (LR_{hi}) and English (LR_{en}):

LR for Hindi:

$$\mathrm{LR}_{hi} = \frac{1}{3}\left(\frac{1}{1} + \frac{1}{3} + \frac{1}{2}\right) = \frac{1}{3}\left(1 + 0.333 + 0.5\right)$$

$$= \frac{1.833}{3} \approx 0.611.$$

LR for English:

$$\mathrm{LR}_{en} = \frac{1}{3}\left(\frac{1}{2} + \frac{1}{1} + \frac{1}{1}\right) = \frac{1}{3}\left(0.5 + 1 + 1\right)$$

$$= \frac{2.5}{3} \approx 0.833.$$

The model's LR scores are approximately 0.611 for Hindi and 0.833 for English, indicating the model detects more English than Hindi in these utterances. This example demonstrates how LR captures the model's ability to rank the relevant languages in code-switched utterances.

D Ground Truth $\mathrm{LR_x}$ Computation

Except for Mandarin-Chinese, where character counts are used, all LangRank ($\mathrm{LR_x}$) computations are based on word counts. For instance, in a Hindi-English code-switched utterance containing 80% Hindi and 20% English words, Hindi is assigned rank 1 and English rank 2. These ranks are computed for every utterance, and the overall $\mathrm{LR_x}$ and $\mathrm{LR_{en}}$ values are then derived using the LangRank formulation described in Section 4.2.2.

E Precision, Recall and F1 Score

E.1 Explanation of Metrics

We report precision, recall, and F1-score as functions of the threshold applied to the probability distribution logits of the languages. The thresholds decide whether a language is relevant to the speech instance or not. Specifically, precision (P), recall (R), and F1-score ($F1$) are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{3}$$

where TP represents true positives, FP denotes false positives, and FN indicates false negatives. By analyzing these metrics, we gain insights into the performance of our model is affected when thresholds are varied.

E.2 Precision, Recall and F1-Score

The predicted set is determined based on a threshold applied to the logits. In Table 4, we report the precision, recall, and F1 scores for the code-switched datasets as a function of the threshold on the logits of each language. Precision increases as the threshold is raised, but eventually starts to decrease. On the other hand, recall consistently decreases as fewer languages are identified when the threshold is increased. As a result, the F1-score shows a consistent decline as the threshold increases.

F KL Divergence for Code-Switched Finetuning

Due to the presence of multiple languages within single utterances in code-switched finetuning, we replaced the cross-entropy loss with Kullback-Leibler (KL) divergence loss:

$$D_{KL}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

where P represents the true probability distribution of the labels and Q represents the predicted probability distribution.

G Ranking of Languages in Code-Switched Datasets

Table 13 presents the language rankings in the ZAEBUC(Habash and Palfreyman, 2022) and the ASCEND datasets (Lovenia et al., 2022). In the ZAEBUC dataset, which contains code-switched Arabic and English, we observe the expected presence of both English and Arabic, alongside unexpected occurrences of Somali and Welsh. Given that this dataset incorporates various dialects and accents, we hypothesize that Somali appears in the predictions due to accent confusion.

In the ASCEND dataset, which features code-switched Mandarin Chinese and English, we note the emergence of two other languages from the CJK family—Japanese and Korean—along with Tibetan. Moreover, Japanese, Chinese, and Korean share historical cultural exchanges and vocabulary borrowing, while Tibetan's connection to Chinese stems from political and religious interactions.

Interestingly, Welsh ranks unusually high in the predictions for both datasets, appearing immediately after the two primary languages. This raises questions, as Welsh is not linguistically related to either Arabic or Mandarin.

H Mismatched accents.

We trained three separate models using 80 examples from Hindi, Bengali, and US-accented English to determine if knowledge from one accent can enhance language identification in other code-switched datasets. Table 5 shows a significant performance boost only when the training accent matches or is similar to the test accent, for both Hindi and Bengali. Models trained with Hindi or Bengali-accented English perform similarly on both the Hindi-English and Bengali-English datasets. This is likely due to shared phonetic and syntactic features between the two languages, which means that the accent-specific features learned during finetuning transfer well across both datasets. However, finetuning with US-accented English did not yield any significant improvement over the baseline, suggesting that mismatched accents do not provide substantial benefits for language identification in this context.

I Influence of Accent Granularity.

We compare models finetuned on two datasets of Indian-accented English: NISP (Kalluri et al., 2020), with Hindi-accented English, and MCV v13 (Ardila et al., 2020) with broader Indian-accented English. As shown in Table 6, finetuning on NISP yields clearer gains in LID performance on code-switched speech, demonstrating that finer accent distinctions enhance performance, especially with small training sets of 80 examples. With 200 examples, performance becomes more comparable across both datasets; however, models finetuned on MCV show increased overfitting on English with high LR_{en} values even for monolingual matrix-language utterances. Notably, our results also indicate that using a broader accent family (e.g., Indian-accented English) can still improve performance when exact accent-matched data is unavailable.

J Performance on accented English.

We finetuned the MMS LID model using LoRA on 80 examples from NISP-accented English. The baseline and finetuned models were then tested on English samples from both the NISP and MCV v13 datasets. As shown in Table 7, the baseline LangRank (LR_{en}) score on NISP English is initially low. However, finetuning the model with Indian English from NISP results in significant improvements in LR_{en}, not only for the NISP dataset but

Threshold on logits	hi-en			bn-en			ar-en			zh-en		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
0.1	0.77	0.59	0.67	0.18	0.18	0.18	0.58	0.45	0.50	0.60	0.54	0.57
0.2	0.80	0.55	0.65	0.18	0.16	0.17	0.58	0.42	0.49	0.61	0.50	0.55
0.3	0.81	0.52	0.63	0.18	0.14	0.16	0.58	0.40	0.47	0.60	0.48	0.53
0.4	0.81	0.50	0.62	0.17	0.13	0.15	0.57	0.39	0.47	0.59	0.47	0.52
0.5	0.80	0.48	0.60	0.16	0.12	0.14	0.56	0.38	0.45	0.57	0.45	0.50

Table 4: Precision, Recall and F1 based on thresholds on logits on code-switched datasets.

English Accent for finetuning	hi-en			bn-en		
	EM	LR_X	LR_{en}	EM	LR_X	LR_{en}
Baseline (No Accent)	509	0.90	0.05	60	0.90	0.05
hi (Hindi Accent)	915	0.85	0.31	331	0.89	0.21
bn (Bengali Accent)	864	0.88	0.28	401	0.87	0.26
us (American Accent)	514	0.90	0.06	75	0.89	0.06

Table 5: Impact of incorporating different English accents used for finetuning the MMS-LID model on code-switched LID performance for Hindi-English (hi-en) and Bengali-English (bn-en) datasets.

English Accent for finetuning	hi-en			bn-en		
	EM	LR_{hi}	LR_{en}	EM	LR_{bn}	LR_{en}
MCV (80 samples)	668	0.89	0.19	1699	0.92	0.14
NISP (80 samples)	915	0.85	0.31	1649	0.90	0.18
MCV (200 samples)	1656	0.81	0.50	1543	0.87	0.45
NISP (200 samples)	1688	0.78	0.53	1600	0.89	0.37

Table 6: Comparison of Indian-accented English from NISP and Mozilla Common Voice on Hindi-English code-switched and monolingual Hindi datasets. The *English Accent* refers to the dataset name and the number of speech samples used for finetuning the MMS-LID model, highlighting the specific accent of English incorporated during training.

also for the MCV English dataset, compared to the baseline performance.

System	LR_{en} (NISP en)	LR_{en} (MCV en)
Baseline MMS-LID	0.47	0.76
MMS-LID(LoRA)	0.99	0.96

Table 7: Comparison of LR_{en} on Indian-accented English from Baseline vs. finetuned models.

K Impact of Training Data Size

Table 8 presents the results of our finetuning experiments conducted with varying sample sizes: 50, 80, 200, and 400 samples of accented English. Optimal performance was achieved with 80 training samples, where the model demonstrated significant improvements in identifying code-switching between English and Hindi/Bengali, while also enhancing detection for Arabic and Mandarin Chinese code-switched instances. Finetuning with 80 samples effectively mitigated overfitting in monolingual datasets by maintaining a lower LR for English, which indicates that the model did not disproportionately favor English in monolingual contexts. These results suggest that 80 training samples strike an ideal balance, offering sufficient exposure to English within code-switched utterances without compromising performance on monolingual samples.

L Incorporating Monolingual Data to Address Overfitting

From Table 8, we observed that finetuning MMS-LID-126 with 200 training samples improved English detection on code-switched datasets but ran into overfitting issues on monolingual samples. To mitigate this and reduce the risk of forgetting non-English languages, we incorporated additional monolingual data from the Fleurs dataset (Conneau et al., 2022) that was used to initially train the MMS-LID-126 model.

We finetuned the model with 200 examples of accented English alongside varying amounts of monolingual data from the corresponding languages (Hindi or Bengali) in the code-switched datasets. As shown in Table 9, although overfitting to English persisted, particularly in monolingual contexts, the inclusion of additional monolingual data led to an increase in exact matches within the code-switched datasets, enhancing the model's ability to recognize and identify code-switching patterns accurately.

M Our Model's Alignment with True Code-Switched and Monolingual Samples

We evaluate our model's performance on a set of monolingual and code-switched utterances and assess how well we are able to differentiate truly monolingual vs. code-switched utterances (based on their ground-truth transcripts). We use the IndicVoices Hindi dataset (Javed et al., 2024) that contains labeled code-switched Hindi-English and monolingual Hindi speech. We compute code-mixed index (CMI) scores for each utterance based

	hi-en			hi		bn-en			bn	
Dataset Size	3136			2000		4275			2000	
System	EM	LR_{hi}	LR_{en}	LR_{hi}	LR_{en}	EM	LR_{bn}	LR_{en}	LR_{bn}	LR_{en}
MMS-LID 50 samples	610	0.88	0.16	0.91	0.10	142	0.89	0.12	0.99	0.04
MMS-LID 80 samples	915	0.85	0.31	0.90	0.18	401	0.87	0.26	0.99	0.09
MMS-LID 200 samples	1688	0.78	0.53	0.89	0.37	1961	0.71	0.73	0.97	0.52
MMS-LID 400 samples	1930	0.65	0.71	0.81	0.52	2445	0.63	0.80	0.94	0.55

(a) Performance metrics for Hindi and Bengali language identification systems.

	ar-en			ar		zh-en			zh	
Dataset Size	6033			2000		1315			2000	
System	EM	LR_{ar}	LR_{en}	LR_{ar}	LR_{en}	EM	LR_{zh}	LR_{en}	LR_{zh}	LR_{en}
MMS-LID 50 samples	1577	0.38	0.40	0.97	0.08	504	0.53	0.27	0.99	0.16
MMS-LID 80 samples	1617	0.38	0.45	0.97	0.13	540	0.53	0.31	0.99	0.17
MMS-LID 200 samples	1740	0.37	0.59	0.97	0.25	615	0.52	0.41	0.99	0.22
MMS-LID 400 samples	1781	0.36	0.72	0.97	0.37	795	0.50	0.67	0.99	0.34

(b) Performance metrics for Arabic and Mandarin Chinese language identification systems.

Table 8: Comparison of LangRank (LR) and Exact Match (EM) scores for language identification across multiple language pairs as finetuning dataset size is varied.

System	hi-en			hi	
	EM	LR_{hi}	LR_{en}	LR_{hi}	LR_{en}
MMS-LID-126	1688	0.78	0.53	0.89	0.37
MMS-LID-126 2:1	1804	0.89	0.45	0.95	0.30
MMS-LID-126 1:1	1937	0.90	0.46	0.96	0.32
MMS-LID-126 1:2	2191	0.92	0.48	0.97	0.34

(a) Impact of adding monolingual data to training on mitigating Hindi forgetting.

System	bn-en			bn	
	EM	LR_{bn}	LR_{en}	LR_{bn}	LR_{en}
MMS-LID-126	1961	0.71	0.73	0.97	0.52
MMS-LID-126 2:1	2008	0.70	0.73	0.97	0.53
MMS-LID-126 1:1	2057	0.70	0.74	0.97	0.53
MMS-LID-126 1:2	1955	0.72	0.72	0.97	0.52

(b) Impact of adding monolingual data to training on mitigating Bengali forgetting.

Table 9: Impact of adding monolingual data to training, along with accented English, on mitigating language forgetting for Hindi and Bengali in code-switched language identification tasks. Here, the ratio $x : y$ means x amount of accented English vs. y amount of monolingual data in the matrix language wherein accented English is fixed to 200 samples. The first row has only accented English.

on their transcript that indicates the extent of language mixing in a sample by measuring the proportion of words from different languages within the same segment. We used the top-scoring 100 samples with the highest CMI as our code-switched Hindi-English subset and the least-scoring 100 samples as our monolingual Hindi subset.

To evaluate the model's performance, we analyzed the confusion matrix from our model for the code-switched samples by classifying a sample as code-switched if both Hindi and English are among the top four predicted languages. (Top-four was essential as the model can confuse closely related languages, such as Hindi and Urdu, due to their lexical overlap.) We compared the performance of the LoRA-finetuned model trained on 80 examples against the baseline model. The results show a significant increase in true positives for code-switched samples, rising from 2 to 41 out of 100 (see Tables 10a and 10b). This improvement confirms the model's enhanced ability to identify code-switching instances compared to the baseline. Additionally, the model's performance on monolingual samples exhibits minimal degradation, indicating that its improved capability to identify code-switched samples does not compromise its effectiveness on monolingual samples.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive = 2	False Negative = 98
Actual Negative	False Positive = 0	True Negative = 100

(a) Confusion Matrix for baseline model

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive = 41	False Negative = 59
Actual Negative	False Positive = 3	True Negative = 97

(b) Confusion Matrix for finetuned model

Table 10: Confusion Matrices for code-switching. Predicted Positive/Negative refers to whether the model predicts sentences as code-switched or not, and Actual Positive/Negative are the reference labels.

Effect of Code-Switching Density. The analysis above provides indirect evidence that model performance correlates positively with the degree

of code-switching present in an utterance. By selecting samples based on the Code-Mixed Index (CMI), we observe that the largest gains from finetuning occur for utterances with high CMI scores, i.e., those exhibiting denser and more frequent language alternations. In contrast, utterances with very low CMI, corresponding to predominantly monolingual speech, show minimal change in performance relative to the baseline, indicating that the finetuning primarily improves sensitivity to mixed-language signals rather than altering monolingual predictions.

While CMI captures the overall proportion of mixed-language content, it does not explicitly model structural properties of code-switching such as the number of switch points or their distribution within an utterance. Prior work suggests that such factors can influence recognition difficulty. We therefore view the current CMI-based analysis as a first-order proxy for code-switching density, and leave a more fine-grained characterization using metrics such as switch-point index or span-level alternation patterns to future work.

N LangRank Detailed Tables

Tables 11a and 12 show all the detailed LangRank values which are plotted in Figure 2. To account for long utterances, we additionally evaluate a variant, MMS-LID (Baseline 4s), which splits inputs into 4-second segments and aggregates predictions via majority voting. This approach improves stability for longer recordings but uses a different total sample count, and therefore is excluded from EM comparisons in Table 2. The effect of this segmentation on LangRank performance is minimal.

O Discussion

In analyzing language identification (LID) in code-switched speech, we observed significant discrepancies in LR_{en} for English across different code-switched datasets at the baseline. Notably, the baseline model excelled in Arabic and Mandarin Chinese code-switching. One potential reason could be due to the data composition of the Fleurs dataset (Conneau et al., 2022). The Fleurs dataset includes both the CJK group—Chinese, Japanese, and Cantonese—and the South Asian (SA) group. The MMS LID 126 model incorporates only Chinese and Japanese from the CJK group. In contrast, it includes all languages from the SA group which encompasses all 14 languages, including variants of Punjabi and Hindi. The high lexical overlap and similar accents among a majority of South Asian languages often lead to their higher ranking than English in code-switched contexts, affecting the model's accuracy in identifying English across diverse scenarios.

While English is identified fairly accurately in the Arabic and Mandarin Chinese code-switched datasets, the LR_X for Arabic and Mandarin Chinese are relatively low. The presence of several other languages in our ranked lists (see Table 13) contributes to a reduction in their overall LRs. This observation highlights the complexities of language identification in code-switched contexts, where the inclusion of multiple languages can skew the performance metrics of specific target languages.

A key finding from our experiments is the overall importance of incorporating English representations from various accents into large multilingual models like MMS. Our results indicate that using accented English data can help mitigate biases that favor specific languages during training. This inclusion is crucial for enhancing the robustness of the model, enabling it to handle real-world code-switching scenarios more effectively.

Table 11: LangRank (LR) scores for various MMS-LID configurations, evaluated on both code-switched and monolingual speech datasets. LR_X and LR_{en} refer to the model's LangRank for the matrix language and embedded English, respectively. The **bolded row (LoRA)** represents the best trade-off overall: it minimizes the total absolute deviation from Oracle LR values across *both dataset types* for a particular language, producing(i) LR_X and LR_{en} closest to Oracle in code-switched settings, and (ii) high LR_X with low spurious LR_{en} in monolingual conditions.

System	hi-en		bn-en		ar-en		zh-en	
	LR_{hi}	**LR_{en}**	**LR_{bn}**	**LR_{en}**	**LR_{ar}**	**LR_{en}**	**LR_{zh}**	**LR_{en}**
MMS-LID (Baseline)	0.90	0.05	0.90	0.05	0.39	0.32	0.54	0.19
MMS-LID (Baseline 4s)	0.86	0.05	0.82	0.06	0.45	0.35	0.55	0.19
MMS-LID (Adapters)	0.59	0.80	0.80	0.64	0.35	0.73	0.52	0.67
MMS-LID (LoRA)	**0.85**	**0.31**	**0.87**	**0.26**	**0.38**	**0.45**	**0.53**	**0.31**
Whisper(Baseline)	0.88	0.37	0.83	0.37	0.37	0.83	0.72	0.66
Oracle	0.94	0.56	0.91	0.59	0.55	0.57	0.83	0.67

(a) Code-switched datasets: Hindi-English (hi-en), Bengali-English (bn-en), Arabic-English (ar-en), and Mandarin-English (zh-en).

System	hi		bn		ar		zh	
	LR_{hi}	**LR_{en}**	**LR_{bn}**	**LR_{en}**	**LR_{ar}**	**LR_{en}**	**LR_{zh}**	**LR_{en}**
MMS-LID (Baseline)	0.92	0.03	0.99	0.02	0.97	0.05	0.99	0.11
MMS-LID (Baseline 4s)	0.90	0.03	0.98	0.02	0.97	0.04	0.98	0.10
MMS-LID (Adapters)	0.62	0.78	0.98	0.51	0.97	0.37	0.99	0.25
MMS-LID (LoRA)	**0.90**	**0.18**	**0.99**	**0.09**	**0.97**	**0.13**	**0.99**	**0.17**
Whisper(Baseline)	0.80	0.33	0.95	0.40	0.97	0.45	0.99	0.43
Oracle	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00

(b) Monolingual datasets: Hindi (hi), Bengali (bn), Arabic (ar), and Mandarin (zh).

Dataset for finetuning	hi-en		hi	
	LR_{hi}	**LR_{en}**	**LR_{hi}**	**LR_{en}**
hi-en (Code-switched)	0.84	0.36	0.87	0.29
hi-accented en	0.85	0.31	0.90	0.18

Table 12: Comparison of Language Recognition (LR) scores for the code-switched Hindi-English finetuned model (*hi-en Code-switched*) versus the Hindi-accented English finetuned model (*hi-accented en*) on both the code-switched Hindi-English dataset (hi-en) and the monolingual Hindi dataset (hi). The table shows the recognition performance for the Hindi language (LR_{hi}) and the English language (LR_{en}) for each model and dataset.

Table 13: LangRanks (LRs) for Language Identification on ZAEBUC and ASCEND Datasets

(a) ZAEBUC (Arabic-English code-switched dataset)

Language	**LR_X**
English	0.4516
Arabic	0.3790
Welsh	0.2102
Somali	0.1405

(b) ASCEND (Mandarin Chinese-English code-switched dataset)

Language	**LR_X**
Mandarin Chinese	0.5322
English	0.3131
Welsh	0.2603
Tibetan	0.2005
Korean	0.1929
Japanese	0.1831
Latin	0.1307