Towards Truly Multilingual ASR: Generalizing Code-Switching ASR to Unseen Language Pairs

Gio Paik [*1,2] **Hyunseo Shin** [*3] **Soungmin Lee** [2,4]

Abstract

Automatic Speech Recognition (ASR) has become a key technology for human–AI interaction. However, code-switching ASR (CS-ASR) remains particularly challenging due to the severe scarcity of multilingual CS speech resources across diverse language pairs. Existing approaches primarily improve CS-ASR performance through synthetic CS speech generation or pair-specific fine-tuning on limited bilingual datasets. Nevertheless, these approaches face an inherent scalability limitation, as support for CS must be developed separately for language pairs whose number grows combinatorially with the number of supported languages. In this work, we investigate whether CS capabilities learned from a limited set of seen language pairs can generalize to unseen language pairs through model merging and domain generalization methods. Our experiments show that merged bilingual CS-ASR models modestly generalize to unseen language pairs, suggesting limited transfer of bilingual CS capabilities across language pairs.

1. Introduction

Despite advances in Automatic Speech Recognition (ASR), code-switching (CS)—the alternation of multiple languages within a single utterance—remains a significant challenge for ASR systems. The core issue lies in the scarcity of CS speech data: for a multilingual ASR model supporting N languages, the number of possible language pairs grows quadratically with N, making it practically infeasible to collect CS speech data for every pair. To address this limitation and enable robust recognition of CS speech commonly used by multilingual speakers, it is necessary to develop ASR systems that can generalize CS capabilities to all $\binom{N}{2}$ language pairs using training data from only a limited subset of language pairs.

In this paper, we investigate whether CS capabilities learned from a subset of language pairs can generalize to unseen language pairs, using speech data from four languages: English (EN), Korean (KO), Japanese (JA), and German (DE). Specifically, we examine whether CS capabilities acquired from relatively accessible language pairs–KO-EN, JA-EN, and DE-EN–can transfer to unseen pairs such as KO-JA and KO-DE. To this end, we explore model merging and domain generalization, and construct small-scale KO-JA[1] and KO-DE CS-ASR evaluation datasets.

Our results show that fine-tuning on one language pair yields slight gains on other pairs, while model merging and domain generalization (DG) can further improve recognition on unseen language pairs. However, the gains remain limited, highlighting the need for methods tailored to the characteristics of CS-ASR, rather than a naive application of existing model merging or domain generalization techniques.

Our contributions are threefold:

1. We systematically investigate whether CS-ASR capabilities learned from specific language pairs can generalize to unseen language pairs.
2. We show the limitations of directly applying existing model merging and domain generalization methods to CS-ASR.
3. We construct the first Korean-Japanese and Korean-German CS speech evaluation datasets, and will open-source the Korean-Japanese dataset.

2. Related Works

2.1. Code-Switching Speech Recognition and Datasets

Code-switching ASR research has largely focused on Chinese-English (Shi et al., 2020; Zhou et al., 2025; Li et al.,

[*]Equal contribution [1]Theta One Korea, Seoul, Republic of Korea [2]Kitsch Labs, Seongnam, Republic of Korea [3]Department of Artificial Intelligence, University of Seoul, Seoul, Republic of Korea [4]College of Computing, Georgia Institute of Technology, GA, USA. Correspondence to: Gio Paik <giopaik0@gmail.com>.

ICML 2026 Workshop on Machine Learning for Audio.

[1]The KO-JA evaluation dataset is available at https://huggingface.co/datasets/thetaone-ai/Korean-Japanese-Code-Switching-Speech.

2025; 2022), with comparatively limited coverage of other English-centric pairs such as English-Korean (Paik et al., 2026) and English-Hindi (Dey & Fung, 2014). In contrast, publicly available datasets for non-English language pairs, such as Korean-Japanese or Korean-German, remain virtually nonexistent. To address this limitation, recent studies have attempted to synthesize code-switching speech using TTS systems (Yan et al., 2025; Yu et al., 2023; Sharma et al., 2020) or by concatenating monolingual speech segments (Lee et al., 2025). However, such approaches often generate acoustically unnatural code-switching speech due to limited CS-aware synthesis capability.

Consequently, most CS-ASR studies have focused on improving recognition performance for individual language pairs. Prior work explored bilingual linguistic biases (Chi & Bell, 2022; Liu et al., 2024), language-specialized architectures (Kulkarni et al., 2023; Zhang et al., 2025), and CS text-based adaptation methods (Nguyen & Tran, 2025; Pandey et al., 2025). However, existing approaches still primarily target seen language pairs rather than generalizing code-switching capabilities to unseen pairs.

2.2. Model Merging

Model merging has recently emerged as an efficient alternative to multi-task retraining for combining independently fine-tuned models. Early approaches such as Task Arithmetic (Ilharco et al., 2023) demonstrated that task-specific parameter differences can be linearly combined in weight space to transfer or compose capabilities across tasks. Subsequent studies further explored more robust merging strategies including TIES-Merging (Yadav et al., 2023), which resolves parameter conflicts through sparse sign agreement, and DARE (Yu et al., 2024), which improves merge robustness through random pruning and rescaling.

Recent work has shown that model merging can effectively combine capabilities across diverse low-resource domains, including multilingual language modeling (Tao et al., 2024; Bandarkar et al., 2025; Shin & Hwang, 2026) and multimodal vision-language models (Chen et al., 2025; Wei et al., 2026). In ASR, Ducorroy & Riad (2025) used model merging to improve robustness to out-of-distribution speech, including disordered speech, while Rolland & Abad (2025) applied model merging to child ASR. However, its application to multilingual CS-ASR, remains unexplored.

2.3. Domain Generalization

Domain generalization has become an important research direction for learning robust representations across heterogeneous domains, particularly in computer vision. Recent work reformulate DG from an optimization perspective, using gradient consistency across domains: MLDG (Li et al., 2018) applies meta-learning to optimize updates for unseen domains, Fish (Shi et al., 2021) maximizes gradient agreement, Fishr (Rame et al., 2022) aligns domain-level gradient variances to regularize loss landscapes, and Gradient-Guided Annealing (GGA) (Ballas & Diou, 2025) mitigates domain overfitting via early-stage gradient alignment. However, most existing DG methods have been explored primarily in computer vision, with relatively limited investigation in low-resource ASR settings.

3. Experiments

3.1. Training Setup

We use the widely adopted multilingual ASR model WHISPER-MEDIUM (Radford et al., 2023) as our backbone and investigate whether code-switching capabilities learned from seen English-centric pairs can improve recognition on unseen non-English-centric pairs.

For the seen language pairs, we fine-tune on three bilingual code-switching datasets, AI-Hub, S. Korea for KO-EN, Shinnosuke et al. for JA-EN, and the DE-EN split of Lee et al. (2025), and evaluate on the human-recorded READ split from Yan et al. (2025). Since no publicly available datasets exist for the unseen KO-JA and KO-DE code-switching pairs, we construct our own evaluation sets. For KO-JA, we collect 450 code-switching utterances whose scripts are written, recorded, and verified by authors proficient in both Korean and Japanese. For KO-DE, we translate the English segments of the KO-EN code-switching dataset from Paik et al. (2026) into German using GPT-5.4 mini (OpenAI, 2026), and ask two graduate students proficient in Korean and German to review and record the translated utterances, resulting in 387 speech samples.

Following prior work on multilingual code-switching ASR (Shi et al., 2020; Zhou et al., 2025; Paik et al., 2026), we use **Mixed Error Rate (MER)**, which accounts for language-specific transcription characteristics within a single utterance. Additional experimental details are provided in Appendix A.

3.2. Fine-Tuning with CS Dataset

We first examine a simple fine-tuning baseline, where WHISPER-MEDIUM is fine-tuned on the code-switching speech data from a single seen language pair. The top of Table 1 shows the performance of the pretrained WHISPER-MEDIUM model and its variants fine-tuned on each language-pair dataset. Overall, fine-tuning on one CS dataset improves recognition not only on the corresponding language pair but also, to some extent, on other code-switching pairs. This trend is particularly pronounced for JA-EN, where the pretrained baseline exhibits a substantially higher MER and all fine-tuning configurations yield clear improvements. However, except for DE-EN, where the pretrained model already

Table 1. Mixed Error Rate (MER) on the dataset for each code-switching language pair. Lower is better.

	SEEN				UNSEEN		
	KO-EN	JA-EN	DE-EN	AVG.	KO-DE	KO-JA	AVG.
WHISPER-MEDIUM	0.26	0.56	0.15	0.33	0.39	0.44	0.41
KO-EN FT	0.12	0.23	0.12	0.16	0.35	0.46	0.40
JA-EN FT	0.14	0.28	0.13	0.18	0.38	0.31	0.35
DE-EN FT	0.14	0.31	0.12	0.19	0.38	0.35	0.36
KO-EN + JA-EN + DE-EN FT	0.11	0.38	0.12	0.20	0.40	0.41	0.41
MODEL MERGING							
TASK ARITHMETIC (ILHARCO ET AL., 2023)							
KO-EN + JA-EN	0.20	0.24	0.18	0.20	0.36	0.53	0.45
KO-EN + DE-EN	0.12	0.29	0.16	0.19	0.36	0.40	0.38
JA-EN + DE-EN	0.17	0.24	0.15	0.19	0.47	0.47	0.47
KO-EN + JA-EN + DE-EN	0.73	0.61	0.34	0.56	0.57	0.96	0.77
TIES (YADAV ET AL., 2023)							
KO-EN + JA-EN	0.11	0.20	0.12	0.14	0.34	0.31	0.32
KO-EN + DE-EN	0.11	0.21	0.12	0.15	0.37	0.39	0.38
JA-EN + DE-EN	0.12	0.25	0.12	0.16	0.44	0.36	0.40
KO-EN + JA-EN + DE-EN	0.11	0.20	0.11	0.14	0.37	0.30	0.34
DARE (YU ET AL., 2024)							
KO-EN + JA-EN	0.21	0.24	0.19	0.21	0.37	0.57	0.47
KO-EN + DE-EN	0.12	0.28	0.16	0.19	0.36	0.40	0.38
JA-EN + DE-EN	0.16	0.28	0.15	0.20	0.48	0.47	0.48
KO-EN + JA-EN + DE-EN	0.74	0.58	0.34	0.55	0.58	0.96	0.77
DOMAIN GENERALIZATION							
FISH (SHI ET AL., 2021)	0.11	0.25	0.15	0.17	0.47	0.53	0.50
FISHR (RAME ET AL., 2022)	0.11	0.29	0.13	0.18	0.35	0.31	0.33
GGA-L (BALLAS & DIOU, 2025)	0.11	0.28	0.13	0.17	0.45	0.40	0.42

performs relatively well, fine-tuning on a different language pair does not consistently produce large MER reductions, suggesting that naive pair-specific adaptation alone provides limited cross-pair generalization.

3.3. Merging

To examine whether model merging can generalize CS-ASR capabilities acquired through fine-tuning on seen language pairs, we merge models fine-tuned on KO-EN, JA-EN, and DE-EN using three methods: Task Arithmetic (Ilharco et al., 2023), TIES (Yadav et al., 2023), and DARE (Yu et al., 2024).

Among the merging approaches, TIES consistently demonstrates the most stable behavior across all pairwise merge settings. In particular, the TIES merge of KO-EN and JA-EN models achieves an average MER of 0.14 on the seen bilingual tasks while maintaining competitive performance on unseen language pairs. Similar trends are observed for the KO-EN + DE-EN and JA-EN + DE-EN settings, suggesting that conflict-aware sparse parameter merging can effectively combine language-pair-specific code-switching capabilities without severe interference.

In contrast, Task Arithmetic and DARE exhibit substantial instability, especially in the three-model merge setting. While pairwise merging remains partially effective, directly combining multiple bilingual CS-ASR models through naive parameter arithmetic often leads to severe degradation.

3.4. Domain Generalization

We also evaluate three domain generalization methods, Fish (Shi et al., 2021), Fishr (Rame et al., 2022), and GGA (Ballas & Diou, 2025), by training on the seen language pairs and measuring their performance on unseen language pairs. For GGA, we use GGA-L, a computationally cheaper variant reported to achieve performance comparable to the full GGA method.

Overall, DG-based fine-tuning does not yield meaningful improvements in MER on unseen pairs, with the exception of Fishr. Fishr improves the average MER on unseen pairs by 0.08 compared with fine-tuning on data from all seen language pairs. However, its absolute MER remains above 0.3, indicating that the improvement is still insufficient for robust unseen pair code-switching recognition.

We hypothesize that this limited gain stems from a mismatch between the assumptions of conventional DG methods and the nature of CS-ASR across language pairs. Standard DG

methods typically assume that task-relevant mechanisms are shared across domains while domain-specific variations change. In contrast, code-switching across different language pairs changes not only the domain but also the output distribution itself, as the target language composition varies across pairs. As a result, naively applying general-purpose DG methods may be insufficient to achieve substantial generalization to unseen code-switching language pairs.

4. Fine-Tuned Parameter Analysis

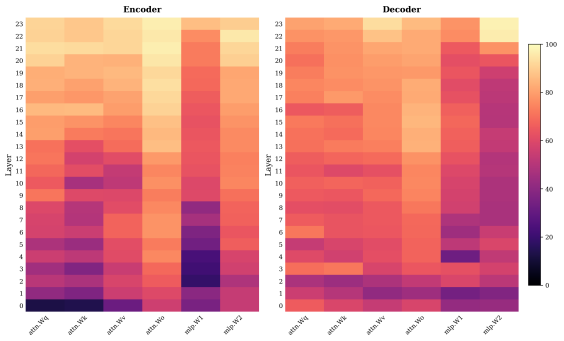

Figure 1. Layer-wise row-level MAV threshold ratios between the pretrained WHISPER-MEDIUM model and the KO-EN code-switching fine-tuned model. Each value represents the percentage of rows whose parameter delta MAV exceeds the predefined threshold.

Figure 1 visualizes the layer-wise row-level Mean Absolute Value (MAV) ratio of parameter deltas between the pretrained WHISPER-MEDIUM model and the KO-EN code-switching fine-tuned model following Bandarkar et al. (2025). We measure the percentage of rows whose delta MAV exceeds a predefined threshold (5×10^{-5}) for each projection matrix in the encoder and decoder layers. Thus, higher values indicate that a larger portion of parameters in the corresponding module were substantially updated during code-switching fine-tuning.

Both the encoder and decoder exhibit progressively larger parameter modifications in higher layers, while lower layers remain relatively stable. This trend suggests that code-switching adaptation primarily occurs in deeper semantic and linguistic representations rather than low-level acoustic processing. The corresponding visualizations for the remaining language pairs are provided in the Appendix B.

5. Limitations

First, performance on unseen language pairs remains limited. Although model merging and domain generalization reduce the average MER on unseen pairs to 0.32, the performance is still far from practical deployment, particularly compared to the sub-0.2 MER achieved after fine-tuning on seen pairs.

Second, both the quantity and diversity of the training and evaluation data are limited. The JA-EN training set contains only 582 utterances from a single speaker, while the KO-JA and KO-DE evaluation sets contain recordings from only two speakers per pair, restricting linguistic and speaker diversity. In addition, our unseen pair experiments involve only combinations of languages already observed during training, and therefore do not evaluate generalization to entirely unseen languages such as French or Chinese. These limitations highlight the need for broader multilingual CS-ASR benchmarks and higher-quality CS speech resources.

Third, our experiments are limited to WHISPER-MEDIUM. A more comprehensive understanding of code-switching generalization will require evaluation across larger Whisper variants and recent audio language models.

Future work should focus on both improving multilingual CS data and developing methods specifically designed for code-switching generalization. Promising directions include analyzing the model components responsible for code-switching behavior, designing domain generalization objectives that explicitly model language-pair shifts, and expanding training and evaluation resources to more diverse language pairs. We view this work as an initial step toward reducing the need to collect code-switching speech data for every possible language pair.

6. Conclusion

In this paper, we investigate whether CS-ASR capabilities learned from a limited set of language pairs can generalize to unseen pairs without requiring pair-specific code-switching data. Using WHISPER-MEDIUM as the backbone, we evaluate fine-tuning and model merging across multilingual code-switching settings involving English, Korean, Japanese, and German.

Our results show that bilingual CS-ASR fine-tuning partially transfers to unseen language pairs, while existing model merging and domain generalization methods remain insufficient to fully bridge the performance gap between seen and unseen pairs. Furthermore, our layer-wise MAV analysis reveals that code-switching adaptation is concentrated in higher encoder and decoder layers, suggesting that generalization to unseen language pairs requires complex task-level adaptations beyond simple domain-level transfer.

These findings highlight the limitations of existing CS-ASR generalization methods and suggest that robust CS-ASR will require architectures and adaptation strategies specifically designed for transferable code-switching capability.

Acknowledgements

This work was supported by the Tech Incubator Program for Startup Korea (RS-2024-00507331) funded by the Ministry of SMEs and Startups (MSS, S. Korea).

References

AI-Hub, S. Korea. Korean-english mixed speech recognition dataset. https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71260.

Ballas, A. and Diou, C. Gradient-guided annealing for domain generalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20558–20568, 2025.

Bandarkar, L., Muller, B., Yuvraj, P., Hou, R., Singhal, N., Lv, H., and Liu, B. Layer swapping for zero-shot cross-lingual transfer in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vQhn4wrQ6j.

Chen, S., Zhang, J., Zhu, T., Liu, W., Gao, S., Xiong, M., Li, M., and He, J. Bring reason to vision: Understanding perception and reasoning through model merging. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9803–9817. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chen25cm.html.

Chi, J. and Bell, P. Improving code-switched ASR with linguistic information. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H. (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 7171–7176, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.627/.

Dey, A. and Fung, P. A Hindi-English code-switching corpus. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL https://aclanthology.org/L14-1705/.

Ducorroy, A. and Riad, R. Robust fine-tuning of speech recognition models via model merging: application to disordered speech. In *Proc. Interspeech 2025*, pp. 3279–3283, 2025.

Goddard, C., Siriwardhana, S., Ehghaghi, M., Meyers, L., Karpukhin, V., Benedict, B., McQuade, M., and Solawetz, J. Arcee's MergeKit: A toolkit for merging large language models. In Dernoncourt, F., Preoţiuc-Pietro, D., and Shimorina, A. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 477–485, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.36. URL https://aclanthology.org/2024.emnlp-industry.36.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Kulkarni, A., Kulkarni, A., Couceiro, M., and Aldarmaki, H. Adapting the adapters for code-switching in multilingual asr. *arXiv preprint arXiv:2310.07423*, 2023.

Lee, S., Chung, W., Um, S., and Kang, H.-G. UniCoM: A universal code-switching speech generator. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13273–13288, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.715. URL https://aclanthology.org/2025.findings-emnlp.715/.

Li, C., Deng, S., Wang, Y., Wang, G., Gong, Y., Chen, C., and Bai, J. TALCS: An open-source Mandarin-English code-switching corpus and a speech recognition baseline. In *Interspeech 2022*, pp. 1741–1745, 2022. URL https://arxiv.org/abs/2206.13135.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Li, Y., Wei, Z., Yu, H., Zhou, H., and Schuller, B. W. Dotame-cs: Daily oriented text audio-mandarin english-code switching dataset. *arXiv preprint arXiv:2501.12122*, 2025. URL https://arxiv.org/abs/2501.12122.

Liu, H., Garcia, L. P., Zhang, X., Khong, A. W. H., and Khudanpur, S. Enhancing code-switching speech recognition

with interactive language biases. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10886–10890, 2024. doi: 10.1109/ICASSP48485.2024.10448335.

Nguyen, T. and Tran, H.-D. Asyncswitch: Asynchronous text-speech adaptation for code-switched asr. *arXiv preprint arXiv:2506.14190*, 2025.

OpenAI. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026.

Paik, G., Kim, Y., Lee, S., Ahn, S., and Kim, C. W. HiKE: Hierarchical evaluation framework for Korean-English code-switching speech recognition. In Demberg, V., Inui, K., and Marquez, L. (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 673–681, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.33. URL https://aclanthology.org/2026.findings-eacl.33/.

Pandey, A., Kumar, K., and Tang, R. Whistle: Deeply supervised, text-only domain adaptation for pretrained speech recognition transformers. *arXiv preprint arXiv:2509.10452*, 2025.

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Rame, A., Dancette, C., and Cord, M. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 18347–18377. PMLR, 2022.

Rolland, T. and Abad, A. Group-aware partial model merging for children's automatic speech recognition. *arXiv preprint arXiv:2511.23098*, 2025.

Sharma, Y., Abraham, B., Taneja, K., and Jyothi, P. Improving Low Resource Code-switched ASR using Augmented Code-switched TTS. In *Interspeech 2020*, 2020. URL https://arxiv.org/abs/2010.05549.

Shi, X., Feng, Q., and Xie, L. The asru 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results. *arXiv preprint arXiv:2007.05916*, 2020. URL https://arxiv.org/abs/2007.05916.

Shi, Y., Seely, J., Torr, P. H. S., Siddharth, N., Hannun, A., Usunier, N., and Synnaeve, G. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Shin, H. and Hwang, W. Layer-wise swapping for generalizable multilingual safety. In Demberg, V., Inui, K., and Marquez, L. (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2223–2238, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.98. URL https://aclanthology.org/2026.eacl-long.98/.

Shinnosuke, T., Yoshifumi, N., Ai, M., Takaaki, S., and Hiroshi, S. Jecs: Japanese-english code-switching speech corpus. https://sites.google.com/site/shinnosuketakamichi/research-topics/jecs_corpus.

Tao, M., Zhang, C., Huang, Q., Ma, T., Huang, S., Zhao, D., and Feng, Y. Unlocking the potential of model merging for low-resource languages. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8705–8720, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.508. URL https://aclanthology.org/2024.findings-emnlp.508/.

Wei, Y., Cheng, R., Jin, W., Yang, E., Shen, L., Hou, L., Du, S., Yuan, C., Cao, X., and Tao, D. Optmerge: Unifying multimodal LLM capabilities and modalities via model merging. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Me0n0iESJY.

Yadav, P., Tam, D., Choshen, L., Raffel, C. A., and Bansal, M. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.

Yan, B., Hamed, I., Shimizu, S., Lodagala, V. S., Chen, W., Iakovenko, O., Talafha, B., Hussein, A., Polok, A., Chang, K., et al. Cs-fleurs: A massively multilingual and code-switched speech dataset. In *Proc. Interspeech 2025*, pp. 743–747, 2025.

Yu, H., Hu, Y., Qian, Y., Jin, M., Liu, L., Liu, S., Shi, Y., Qian, Y., Lin, E., and Zeng, M. Code-switching text generation and injection in mandarin-english asr. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. URL https://arxiv.org/abs/2303.10949.

Yu, L., Yu, B., Yu, H., Huang, F., and Li, Y. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

Zhang, F., Geng, W., Huang, H., Shan, Y., Yi, C., and Qu, H. Boosting code-switching asr with mixture of experts enhanced speech-conditioned llm. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Zhou, J., Guo, Y., Zhao, S., Sun, H., Wang, H., He, J., Kong, A., Wang, S., Yang, X., Wang, Y., et al. Cs-dialogue: A 104-hour dataset of spontaneous mandarin-english code-switching dialogues for speech recognition. *arXiv preprint arXiv:2502.18913*, 2025. URL https://arxiv.org/abs/2502.18913.

A. Experimental Details

Training We adopt WHISPER-MEDIUM (Radford et al., 2023) as the backbone model. All fine-tuning experiments on a single language pair are performed with a batch size of 8 for 73 training steps. For KO-EN + JA-EN + DE-EN FT and domain generalization experiments, we use a batch size of 9 for 195 training steps. We employ the AdamW optimizer with a cosine learning rate decay schedule and a linear warmup phase corresponding to 10% of the total training steps. For model merging methods, including Task Arithmetic, TIES, and DARE, we use MergeKit (Goddard et al., 2024). All experiments are conducted using PyTorch 2.8.0 on NVIDIA GeForce RTX 4090 GPUs.

B. Parameter Analysis

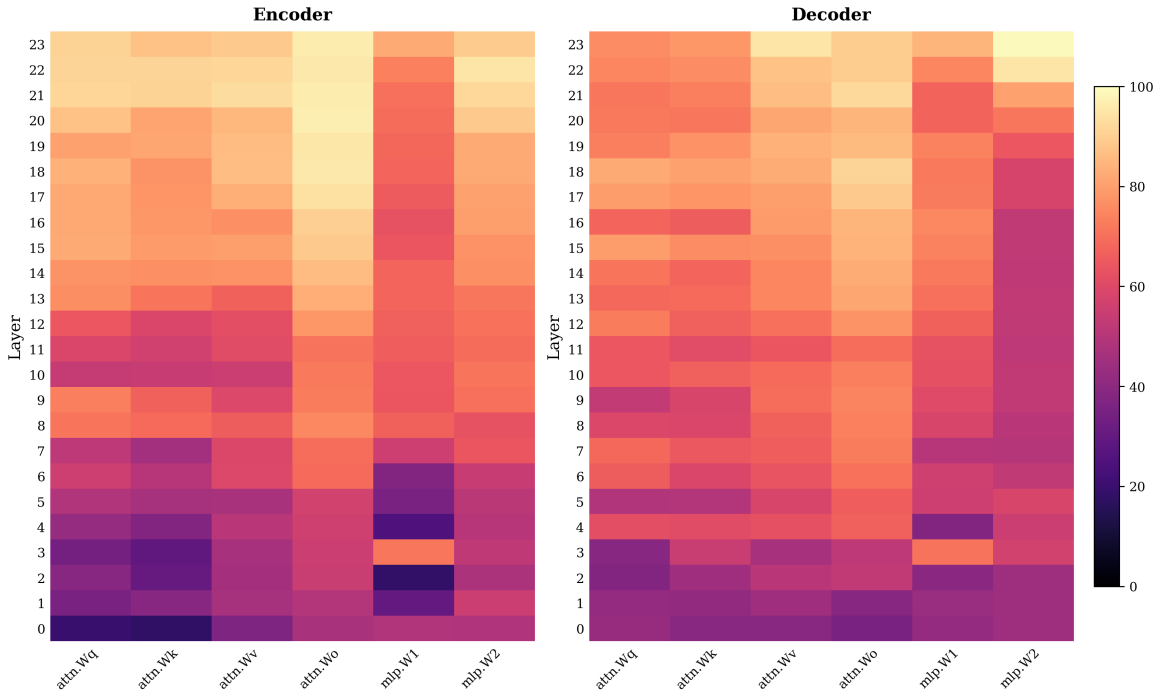

Figure 2. Layer-wise row-level MAV threshold ratios between the pretrained WHISPER-MEDIUM model and the JA-EN code-switching fine-tuned model.

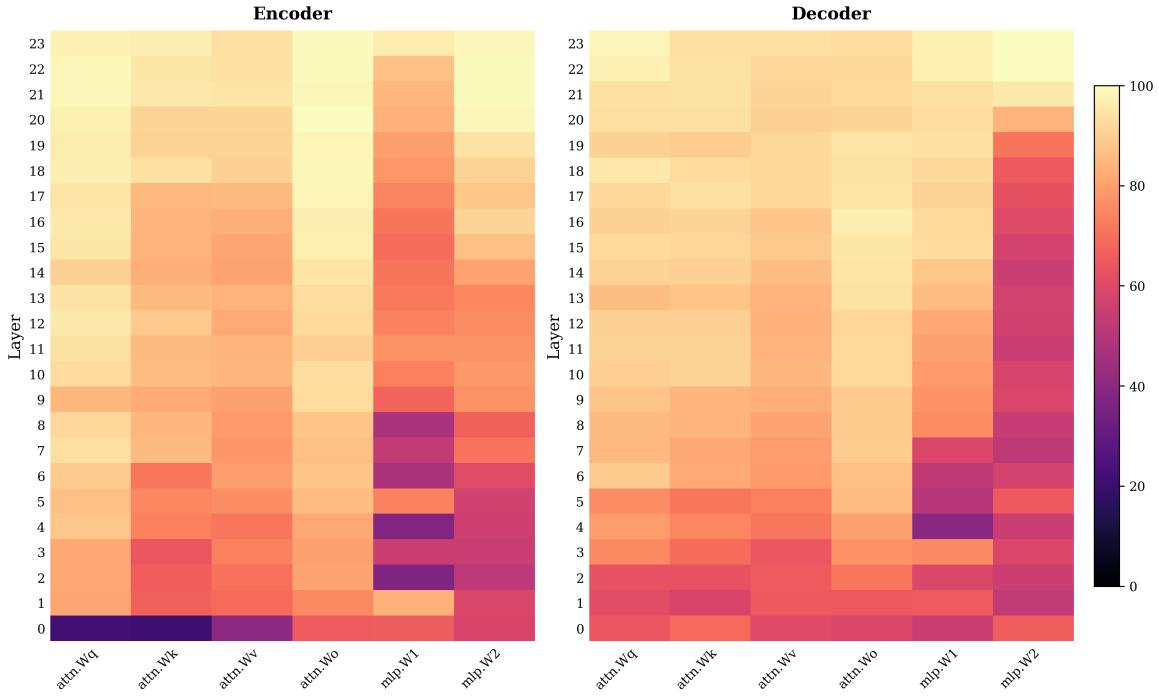

Figure 3. Layer-wise row-level MAV threshold ratios between the pretrained WHISPER-MEDIUM model and the DE-EN code-switching fine-tuned model.