HiKE: Hierarchical Evaluation Framework for Korean-English Code-Switching Speech Recognition

Gio Paik[1,5,*], Yongbeom Kim[2,5], Soungmin Lee[3,5], Sangmin Ahn[1,2,5,†], Chanwoo Kim[1,4,5,†]

[1]Theta One AI, [2]Seoul National University, [3]Georgia Institute of Technology,
[4]Williams College, [5]ROKAF Reserve Forces

***Corresponding Author:** giopaik0@gmail.com †: Equal Contribution

Abstract

Despite advances in multilingual automatic speech recognition (ASR), code-switching (CS), the mixing of languages within an utterance common in daily speech, remains a severely underexplored challenge. In this paper, we introduce HiKE: the Hierarchical Korean-English code-switching benchmark, the first globally accessible non-synthetic evaluation framework for Korean-English CS, aiming to provide a means for the precise evaluation of multilingual ASR models and to foster research in the field. The proposed framework not only consists of high-quality, natural CS data across various topics, but also provides meticulous loanword labels and a hierarchical CS-level labeling scheme (word, phrase, and sentence) that together enable a systematic evaluation of a model's ability to handle each distinct level of code-switching. Through evaluations of diverse multilingual ASR models and fine-tuning experiments, this paper demonstrates that although most multilingual ASR models initially exhibit inadequate CS-ASR performance, this capability can be enabled through fine-tuning with synthetic CS data. HiKE is available at https://github.com/ThetaOne-AI/HiKE.

1 Introduction

Recent advances in Automatic Speech Recognition (ASR) (Radford et al., 2023; Puvvada et al., 2024; Saon et al., 2025) have pushed error rates below 5% on standard monolingual ASR benchmarks (Panayotov et al., 2015; Conneau et al., 2022; Rousseau et al., 2012), enabling novel applications such as vibe coding, AI-assisted language education and automated podcast summarization. These advancements are fundamentally redefining human-computer interaction. However, it remains a significantly underexplored question whether the performance of these ASR models in monolingual settings can be extended to Code-Switching (CS) scenarios, where multiple languages are mixed within a single utterance. Consequently, the field of CS-ASR remains underdeveloped (Agro et al., 2025), especially for language pairs involving low-resource and typologically distant languages such as Korean and English. This research gap significantly impairs the user experience for the large global population of multilingual individuals who use CS as a natural, everyday part of communication, particularly in regions where English is not the primary language.

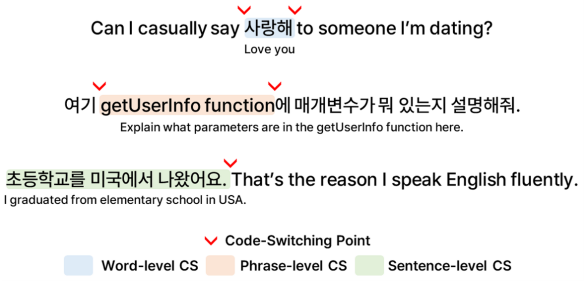

Figure 1: **Code-Switching Examples by CS-Level**

To address this research gap, this paper introduces **HiKE**: **Hi**erarchical **K**orean-**E**nglish code-switching benchmark consisting of 1,121 high-quality CS utterances covering various topics (e.g., Software Engineering, Language Education). To reflect the various forms of CS that occur in real-world scenarios, we labeled the utterances according to three hierarchical CS-levels (i.e., word-, phrase- and sentence-level), as illustrated in Figure 1. Using HiKE, we conduct a two-part analysis. First, we evaluate ten multilingual ASR models, spanning a range of architectures and model sizes, to assess their Korean-English CS-ASR capabilities. Second, we compare the CS-ASR performance achieved by fine-tuning with two different data types: natural word- and phrase-level CS data and synthesized sentence-level CS data.

The contributions of this paper are three-fold: **First,** to the best of our knowledge, HiKE is the first to release a publicly available non-synthetic

Findings of the Association for Computational Linguistics: EACL 2026, pages 673–681

Topic	CS-Level			Total	Proportion
	word	phrase	sentence		
academic	40	110	7	157	14.0%
business	45	112	12	169	15.1%
entertainment	53	29	8	90	8.0%
everyday conversation	69	76	13	158	14.1%
language education	82	75	1	158	14.1%
medical	28	48	0	76	6.8%
software development	43	113	6	162	14.5%
travel and culture	97	44	10	151	13.5%
Total	457	607	57	**1,121**	**100%**

Table 1: **Number of Utterances by Topic & CS-Level**

Korean-English CS speech recognition benchmark, which includes loanword labels and CS-level annotations. **Second,** leveraging our loanword and hierarchical CS-level annotations, we precisely measure how the performance of 10 multilingual ASR models varies depending on the type of CS. **Third,** our fine-tuning experiments demonstrate that a model's CS-ASR capabilities can be effectively enabled through fine-tuning, and that this is achievable not only with natural CS data but also with synthetic data created by concatenating monolingual utterances.

2 Related Work

2.1 Automatic Speech Recognition

Early deep learning-based ASR systems were often streaming-based, using architectures such as Connectionist Temporal Classification (CTC) to make predictions on short audio frames (Baevski et al., 2020; Hsu et al., 2021; An et al., 2024). A subsequent paradigm shift occurred towards non-streaming, Transformer-based encoder-decoder models (Vaswani et al., 2017; Radford et al., 2023; Barrault et al., 2023), which leverage full audio context to achieve superior accuracy and robustness. The most recent trend involves directly leveraging pretrained large language models (LLMs) (Gemma Team, 2025; Goel et al., 2025; Hurst et al., 2024), offering high accuracy and task extensibility, but requiring immense computational resources.

2.2 Code-Switching Speech Datasets

Publicly available CS-ASR datasets are exceptionally rare due to the inherent challenges of data collection. Even for Mandarin-English, whose large speaker population has resulted in a relative abundance of data (Shi et al., 2020; Zhou et al., 2025; Li et al., 2025), only a handful of public benchmarks exist. The scarcity of datasets is particularly acute for typologically distant language pairs like Korean-English, presenting a critical bottleneck for research into improving CS-ASR performance. While a government-funded Korean dataset (AI-Hub, S. Korea) exists, its use is restricted to Korean nationals, making it inaccessible to the global research community.

To address the scarcity of Code-Switching data, previous research has explored methods such as concatenating monolingual utterances to generate CS samples (Hussein et al., 2024; Seki et al., 2018) or utilizing Text-to-Speech (TTS) models to produce synthetic CS data (Yan et al., 2025; Yu et al., 2023; Sharma et al., 2020) for the training and evaluation of Automatic Speech Recognition (ASR) systems. However, concatenated data often suffers from mid-utterance discontinuities in vocal traits (e.g., gender and recording environment) and is restricted to sentence-level CS. Similarly, TTS-generated data may exhibit poor audio quality and, due to its synthetic nature, fails to capture the diverse CS patterns and acoustic variability found in real-world environments, making it unsuitable for evaluating performance in authentic scenarios. To the best of our knowledge, HiKE is the first globally accessible non-synthetic Korean-English CS benchmark.

3 HiKE

3.1 Script Writing & Cloning

We built our dataset through a human-LLM collaborative process to ensure both high quality and minimal human effort. We began by manually authoring 575 seed scripts across 8 topics (Table 1). Each script was then used as a one-shot example to prompt CLAUDE-3.5-SONNET (Anthropic, 2024) to generate a new script mimicking the original's topic and CS structure, utilizing the prompt detailed in Appendix D. To prevent excessive similarity among the generated scripts and maintain a balance between data volume and diversity, we opted to generate only one clone per original script. As a final quality control step, all generated scripts were manually reviewed and corrected by the authors.

3.2 Recording

We recruited 13 bilingual Korean-English speakers, each of whom recorded 50 to 100 scripts in a quiet environment using a web-based interface on their personal devices, such as laptops and mobile devices. This process yielded an initial batch

	# Params	Mixed Error Rate (MER)				Point of Interest Error Rate(PIER)				Monolingual	
		Word	Phrase	Sentence	Overall	Word	Phrase	Sentence	Overall	KOR	ENG
SENSEVOICE-SMALL (An et al., 2024)	234M	27.4	36.4	**23.6**	32.5	52.2	56.5	**37.1**	54.7	6.4	7.6
WHISPER-TINY (Radford et al., 2023)	38M	73.7	111.4	**36.2**	93.6	81.7	77.9	**66.9**	78.9	11.9	14.6
WHISPER-BASE (Radford et al., 2023)	74M	116.5	81.8	**25.4**	91.2	90.2	73.2	**38.7**	78.1	7.8	9.8
WHISPER-SMALL (Radford et al., 2023)	244M	50.6	41.3	**19.9**	43.5	58.0	46.7	**31.5**	50.1	4.5	8.3
WHISPER-MEDIUM (Radford et al., 2023)	769M	39.1	27.9	**16.8**	31.3	41.9	41.3	**30.6**	41.3	3.4	4.6
WHISPER-LARGE (Radford et al., 2023)	1.5B	28.5	25.3	**20.0**	26.1	45.6	31.3	**25.8**	36.0	3.2	4.4
SEAMLESS-M4T-V2-LARGE (Barrault et al., 2023)	2.3B	108.8	70.1	**61.0**	83.6	72.4	60.9	**57.3**	64.7	6.4	6.3
GEMMA-3N (Gemma Team, 2025)	8B	99.6	64.6	**54.3**	76.6	69.2	47.8	**40.3**	54.8	10.7	13.0
AUDIO-FLAMINGO-3 (Goel et al., 2025)	8.3B	**78.5**	82.2	79.4	80.7	104.1	**97.8**	112.9	100.2	25.1	8.8
GPT-4O-TRANSCRIBE (Hurst et al., 2024)	N/A	**15.6**	25.0	28.3	21.8	**25.8**	30.2	32.3	28.8	2.5	3.3

Table 2: **Benchmark Results.** For each model, the **best** and worst scores are bolded and underlined, respectively. Monolingual performance is measured on the FLEURS dataset, using CER for Korean and WER for English.

of 1,150 recordings. Following a manual review by the authors, 29 samples that deviated from the script were discarded, resulting in a final curated dataset of 1,121 high-quality utterances totaling approximately 2.2 hours.

3.3 Metrics

Given that CS-ASR involves mixing languages with different linguistic properties, many prior works (Shi et al., 2020; Zhou et al., 2025; Li et al., 2025) have adopted the Mixed Error Rate (MER), which evaluates character-based languages such as Mandarin and Korean at the character level and word-based languages such as English at the word level. In contrast to MER, which assesses the entire utterance, the Point of Interest Error Rate (PIER) (Ugan et al., 2025) was later proposed to specifically evaluate performance at the points where language transitions occur. In this paper, we evaluate the CS-ASR capabilities of models using both MER and PIER. For our PIER evaluation, we tagged the words at the locations where code-switching occurs. This allowed us to assess whether the ASR model can accurately switch languages precisely at these transition points.

3.4 Hierarchical CS-Level

In contrast to previous work (Shi et al., 2020; Lovenia et al., 2022; Li et al., 2022) that only classified CS by its location within the utterance (inter- vs. intra-sentential), we propose a more granular, three-level classification to analyze how models handle intra-sentential CS between typologically distant languages. HiKE categorizes code-switching into three distinct levels: sentence, word, and phrase. Sentence-level CS is the most predictable, as switches occur only at utterance boundaries. Word-level CS involves the substitution of single lexical units, primarily testing a model's bilingual lexicon. Phrase-level CS, in contrast, poses a more complex grammatical challenge, as it can introduce irregular structures like altered word order, a difficulty that is significantly amplified for typologically distant pairs such as Korean and English.

3.5 Loanwords Post-processing

Loanwords are words adopted from a foreign language and adapted to the phonology and orthography of the new language. For example, the Korean loanword '버스' [bɔs] and the English word 'bus' [bʌs] are pronounced almost identically and can be used interchangeably in a CS context. This creates an evaluation challenge: If a ground-truth label is strictly constrained to either Hangul or the Roman alphabet, a model producing the alternate—yet perfectly valid—transcription would be unfairly penalized. This ambiguity introduces significant noise into the assessment of CS performance. To avoid this problem, we meticulously labeled all loanwords contained in our dataset. Subsequently, during evaluation, both the Korean and English versions of these loanwords were treated as valid answers. Through our loanword labeling process, we achieved a more precise evaluation by decreasing measurement noise by an average of 4.9% in MER and 7.9% in PIER.

4 Experiments

4.1 CS-ASR of Multilingual ASR Models

We evaluated 10 multilingual ASR models with diverse architectures (CTC, Transformer, and LLM-based), assessing their CS-ASR performance using MER and PIER metrics. As shown in Table 2, a severe performance drop occurred on CS data across all models, with the MER increasing by a factor of 3 to 13 compared to monolingual perfor-

	Mixed Error Rate (MER)				Point of Interest Error Rate (PIER)				Monolingual	
	Word	Phrase	Sentence	Overall	Word	Phrase	Sentence	Overall	KOR	ENG
WHISPER-MEDIUM	39.1	27.9	16.8	31.3	41.9	41.3	30.6	41.3	3.4	4.6
(a) FT with Natural Intra-Sentential CS Data	**8.3**(-30.8)	**9.6**(-18.4)	7.4(-9.4)	**9.0**(-22.3)	**18.6**(-23.3)	**19.9**(-21.4)	21.0(-9.7)	**19.5**(-21.8)	6.0(+2.6)	5.2(+0.6)
(b) FT with Synthetic Inter-sentential CS Data	19.1(-20.0)	25.6(-2.3)	5.8(-11.0)	22.1(-9.2)	33.7(-8.2)	35.6(-5.7)	14.5(-16.1)	34.5(-6.8)	3.7(+0.3)	5.1(+0.5)
(a) + (b) FT with Both Data	20.7(-18.4)	21.7(-6.3)	**4.8**(-11.9)	20.4(-11.0)	27.3(-14.6)	37.6(-3.7)	**11.3**(-19.4)	33.6(-7.7)	3.9(+0.5)	4.9(+0.3)

Table 3: **Fine-Tuning (FT) Results.** For each metric, the **best** scores are highlighted in bold. Monolingual performance is measured on the FLEURS dataset, using CER for Korean and WER for English.

mance, revealing significant limitations for practical use. A closer look at specific architectures reveals further nuances. For instance, the CTC-based SENSEVOICE-SMALL, despite showing stable performance across CS-levels and a better overall MER than the similarly-sized WHISPER-SMALL, still exhibited a high error rate at the actual code-switching points. This suggests that while CTC architectures may be robust in overall transcription, they struggle specifically with the precise moment of language transition.

In contrast, models leveraging Large Language Models (LLMs) trained on extensive text corpora displayed distinct behavioral patterns compared to speech-centric models. Specifically, despite having over five times the parameter count of WHISPER-LARGE, GEMMA-3N and AUDIO-FLAMINGO-3 failed to achieve competitive results in CS-ASR, with both MER and PIER exceeding 54. GPT-4O-TRANSCRIBE was the only LLM-based model to outperform WHISPER-LARGE. Notably, while most non-LLM models performed best on sentence-level CS and worst on word-level CS, GPT-4O-TRANSCRIBE exhibited the opposite trend, achieving its highest performance on word-level CS and its weakest on sentence-level CS. We hypothesize that this reflects the distribution of its training data, as large text corpora are rich in word-level CS but contain comparatively few instances of sentence-level CS.

To analyze the effect of model scale on CS-ASR performance, we evaluated the Whisper family of models. The results show a clear correlation between size and capability: CS-ASR performance, nearly absent in the smallest models (e.g., Tiny and Base), gradually emerges with increasing scale. Despite this trend, even the largest model's error rate on CS data was over five times higher than on monolingual data, indicating that model scaling alone is an insufficient solution for achieving practical CS-ASR performance.

4.2 Fine-Tuning with a Synthetic CS Dataset

4.2.1 Experimental Details

To investigate the effect of fine-tuning with CS-ASR performance, we prepared two distinct types of training data. The first was a natural, intra-sentential (word- and phrase-level) CS dataset from AIHub (AI-Hub, S. Korea). The second was a synthetic sentence-level CS dataset, which we generated by concatenating monolingual Korean and English utterances from the FLEURS (Conneau et al., 2022) and Common Voice (Ardila et al., 2020) datasets. Using these, we fine-tuned WHISPER-MEDIUM (Radford et al., 2023) under three conditions: using the intra-sentential CS set only, the synthetic inter-sentential set only, and a combination of both.

4.2.2 Results

The results in Table 3 reveal that fine-tuning is an effective method for enabling CS-ASR, and that this can be achieved not only with natural intra-sentential CS data **(a)** but also with synthetic inter-sentential CS data **(b)** created by concatenating monolingual utterances. As shown in row **(b)** of Table 3, fine-tuning of synthetic CS data alone improved both the overall MER and PIER by more than 6.8%. Given the wide availability and relative ease of collecting monolingual data compared to authentic CS data, this finding suggests that data synthesis via utterance concatenation represents a promising and cost-effective direction for training CS-ASR models, especially in resource-constrained scenarios. However, fine-tuning on natural intra-sentential CS data **(a)** yielded greater performance improvements than fine-tuning on synthetic inter-sentential data **(b, c)** across almost all metrics; the only exceptions were the MER and PIER for sentence-level CS. Although this result indicates that it is currently preferable to use natural intra-sentential CS data for fine-tuning when available, we anticipate that these limitations can be overcome. With the development of more so-

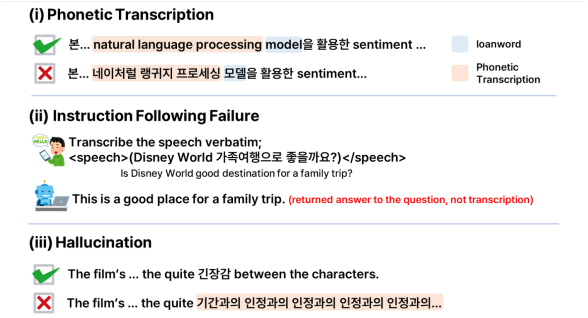

Figure 2: **CS-ASR Fail Cases**

phisticated data synthesis or fine-tuning techniques, it may become possible to train robust CS-ASR models using only synthetic data.

4.3 Qualitative Results

In our analysis of the CS transcription results, we primarily observed three types of errors, as illustrated in Figure 2. **(i)** Phonetic Transcription refers to cases where, for words that are not loanwords, the model does not transcribe them in the correct language but instead writes them out phonetically using the script of the other language. This error was commonly observed across all models. **(ii)** Instruction Following Failure is an error that occurs in multi-task models (e.g., WHISPER) capable of handling tasks beyond transcription, such as translation and question answering. This was especially pronounced with AUDIO-FLAMINGO-3; while this error was rare in monolingual settings, its high frequency in CS environments made the model unreliable for transcription without a separate verification step. Finally, **(iii)** Hallucination is an error common in seq2seq models, including those based on Transformers and LLMs. It refers to cases where the model incorrectly generates repetitive or excessive content that is not present in the audio.

Qualitatively, while errors like phonetic transcription often occurred even with monolingual data, instruction following failures and hallucinations increased markedly in CS data. We attribute this trend to the fact that during training, ASR models are exposed to abundant monolingual data but extremely rare CS data, which hinders the generalization of their performance to CS scenarios.

5 Conclusion

In this paper, we propose HiKE, the first public high-quality hierarchical benchmark for Korean-English CS-ASR with hierarchical CS-level labels and loanword labels. Our evaluation of 10 multilingual models in HiKE shows that strong monolingual performance does not generalize to CS scenarios. Specifically, we found that models are less accurate on word- and phrase-level CS, which feature dense, irregular switch points and complex grammatical structures, in contrast to the more predictable sentence-level CS. Furthermore, our fine-tuning experiments demonstrate that a model's CS-ASR capabilities can be improved, even with synthetic inter-sentential CS data created by concatenating monolingual utterances. We believe that this work will serve as a foundation for future research in CS-ASR, including developing models for diverse language pairs beyond Korean-English, generating high-quality synthetic CS data, and analyzing the generalization of CS-ASR capabilities.

Limitations

Our study has two primary limitations. First, the scarcity of ASR models that support both Korean and English precluded a broad comparative study of diverse architectures. This was especially true for LLM-based models, as only GPT-4O-TRANSCRIBE yielded a meaningful CS-ASR performance, which prevented a reliable analysis of their common characteristics. Second, the scarcity of high-quality CS data restricted our fine-tuning experiments to a small scale, precluding a thorough investigation into methods for eliciting robust CS capabilities, such as by scaling up synthetic data. We believe these limitations point to several avenues for future research.

Acknowledgments

We appreciate Hyunwoo Kim for proofreading and the anonymous reviewers for their insightful feedback and suggestions. Special thanks to DongChan Shin (@DongChanS) for pointing out issues in the preprocessing pipeline on GitHub. We also thank the Theta One Team for supporting our work.

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

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. *arXiv preprint arXiv:2507.07741*.

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

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, and 1 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Anthropic. 2024. Claude 3.5 sonnet model card addendum.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 4218–4222.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

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. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Gemma Team. 2025. Gemma 3n.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, volume 29, pages 3451–3460.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Amir Hussein, Dorsa Zeinali, Ondřej Klejch, Matthew Wiesner, Brian Yan, Shammur Chowdhury, Ahmed Ali, Shinji Watanabe, and Sanjeev Khudanpur. 2024. Speech collage: code-switched audio generation by collaging monolingual corpora. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12006–12010.

Chengfei Li, Shuhao Deng, Yaoping Wang, Guangjing Wang, Yaguang Gong, Changbin Chen, and Jinfeng Bai. 2022. TALCS: An open-source Mandarin-English code-switching corpus and a speech recognition baseline. In *Interspeech 2022*, pages 1741–1745.

Yupei Li, Zifan Wei, Heng Yu, Huichi Zhou, and Björn W Schuller. 2025. Dota-me-cs: Daily oriented text audio-mandarin english-code switching dataset. *arXiv preprint arXiv:2501.12122*.

Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Peng Xu, Yan Xu, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J. Barezi, Qifeng Chen, Xiaojuan Ma, Bertram Shi, and Pascale Fung. 2022. ASCEND: A Spontaneous Chinese-English Dataset for Code-switching in Multi-turn Conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 7259–7268.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Krishna C. Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Less is More: Accurate Speech Recognition & Translation without Web-Scale Data. In *Interspeech 2024*, pages 3964–3968.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning (ICML)*, pages 28492–28518.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2012. TED-LIUM: an Automatic Speech Recognition dedicated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 125–129.

George Saon, Avihu Dekel, Alexander Brooks, Tohru Nagano, Abraham Daniels, Aharon Satt, Ashish Mittal, Brian Kingsbury, David Haws, Edmilson Morais, and 1 others. 2025. Granite-speech: open-source speech-aware llms with strong english asr capabilities. *arXiv preprint arXiv:2505.08699*.

Hiroshi Seki, Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R. Hershey. 2018. An End-to-End Language-Tracking Speech Recognizer for Mixed-Language Speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4919–4923.

Yash Sharma, Basil Abraham, Karan Taneja, and Preethi Jyothi. 2020. Improving Low Resource Code-switched ASR using Augmented Code-switched TTS. In *Interspeech 2020*.

Xian Shi, Qiangze Feng, and Lei Xie. 2020. The asru 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results. *arXiv preprint arXiv:2007.05916*.

Enes Yavuz Ugan, Ngoc-Quan Pham, Leonard Bärmann, and Alex Waibel. 2025. PIER: A Novel Metric for Evaluating What Matters in Code-Switching. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Brian Yan, Injy Hamed, Shuichiro Shimizu, Vasista Lodagala, William Chen, Olga Iakovenko, Bashar Talafha, Amir Hussein, Alexander Polok, Kalvin Chang, and 1 others. 2025. Cs-fleurs: A massively multilingual and code-switched speech dataset. *arXiv preprint arXiv:2509.14161*.

Haibin Yu, Yuxuan Hu, Yao Qian, Ma Jin, Linquan Liu, Shujie Liu, Yu Shi, Yanmin Qian, Edward Lin, and Michael Zeng. 2023. Code-switching text generation and injection in mandarin-english asr. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jiaming Zhou, Yujie Guo, Shiwan Zhao, Haoqin Sun, Hui Wang, Jiabei He, Aobo Kong, Shiyao Wang, Xi Yang, Yequan Wang, and 1 others. 2025. Cs-dialogue: A 104-hour dataset of spontaneous mandarin-english code-switching dialogues for speech recognition. *arXiv preprint arXiv:2502.18913*.

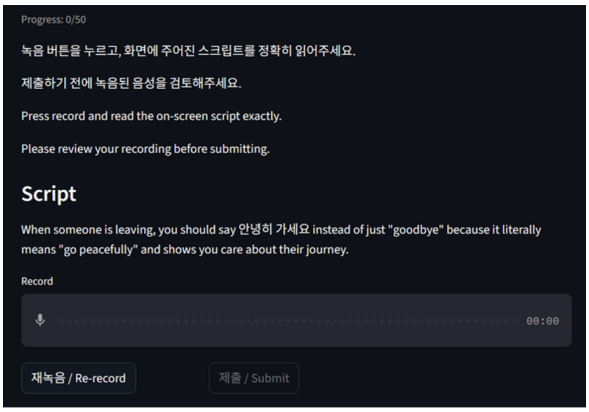

Figure 3: **Screenshot of Recording Tool**

Characteristic	Count
Gender	
Male	11
Female	2
Educational Level	
Undergraduate	6
Bachelor's	7
Total	13

Table 4: **Demographics of Participants**

A Experimental Setup

Unless otherwise specified, all experiment results were obtained from a single run on a NVIDIA RTX 6000 Ada GPU. We used pyTorch 2.8.0 and transformers 4.56.2 for our experiments.

In finetuning experiments, all models were finetuned for approximately 1 epoch on a single A100 GPU with a batch size of 16, using a cosine annealing scheduler with a 10% warmup and a peak learning rate of $1e-5$.

B Dataset Recording

As mentioned in Section 3.2, participants performed the data recordings using a web-based interface like the one shown in Figure 3. The webpage provided recording instructions and scripts in both Korean and English, and allowed participants to review their own recordings and perform re-takes if necessary. The demographic distribution of the participants, specifically regarding gender and educational background, is summarized in Table 4.

Referring to the example below, create a new Korean-English code-switching sample. First, analyze the example to identify its characteristics and themes, then return a new code-switching sample that reflects those characteristics, wrapped in a <sample> tag.
<example>
{example}
</example>

Figure 4: **Script Cloning Prompt**

C License

Our experiments utilize the official code implementation of PIER (Ugan et al., 2025) (Apache 2.0 License), along with other standard machine learning libraries. Our own evaluation code, developed for the HiKE benchmark, will also be publicly released under the Apache 2.0 License.

D Script Cloning Prompt

Figure 4 shows the prompt used to clone a new script based on a given original script.

E CS-Level Determination

To ensure consistent and automated determination of CS-levels, we adopted the following hierarchical procedure. First, a sample is classified as sentence-level code-switching if it contains at least one sentence composed entirely of English and another entirely of Korean. Among the remaining samples, those containing a sequence of two or more consecutive words in a language different from the rest are categorized as phrase-level code-switching. Finally, any remaining samples where the code-switching consists of only a single, non-consecutive word are classified as word-level code-switching. Figure 5 presents the pseudocode for this process. L_{main} represents the main language of the utterance annotated by the authors, whereas L_{other} indicates the other language.

Algorithm 1 CS-Level Determination Procedure

procedure LEVEL-DETECTING($RawText, L_{main}, L_{other}$)
▷ Step 1 & 2: Pre-processing & Mapping
    $CleanText \leftarrow$ RemoveSpecialChars($RawText$)
    $Seq \leftarrow$ Map characters to $\{L_{main}, L_{other}\}$
    $Segments \leftarrow$ Split(Seq, delimiters $= \{., !, ?\}$)
    ▷ Step 3: Level Checking (Independent Detection)
    for each $segment$ in $Segments$ **do**
        if $segment$ consists **only** of L_{other} **then**
            $has\_sentence \leftarrow$ true
        end if
        if $segment$ contains a sequence of (L_{other}, L_{other}) **then**
            $has\_phrase \leftarrow$ true
        end if
        if $segment$ contains an isolated unit of L_{other} **then**
            $has\_word \leftarrow$ true
        end if
    end for
    ▷ Step 4: Priority-based Classification
    if $has\_sentence$ **then**
        return "sentence"
    else if has_phrase **then**
        return "phrase"
    else if has_word **then**
        return "word"
    else
        return "None"
    end if
end procedure

Figure 5: CS-Level Determination Procedure