

CS-YODAS: A Mined Dataset of In-the-Wild Code-Switched Speech

Brian Yan¹, Qingzheng Wang¹, Matthew Wiesner², Anuj Diwan³, Olga Iakovenko⁴
Alexander Polok⁵, Injy Hamed⁶, Shuichiro Shimizu⁷, Iris Emerman
Thomas Hain⁴, David R. Mortensen¹, Peter Viechnicki, Shinji Watanabe¹

¹Carnegie Mellon University, ²Johns Hopkins University, ³University of Texas at Austin,
⁴University of Sheffield, ⁵Brno University of Technology, ⁶MBZUAI, ⁷Kyoto University
byan@cmu.cs.edu

Abstract

We present CS-YODAS, a Creative Commons-licensed dataset of in-the-wild code-switched speech mined from multilingual YouTube data. Code-switching (CS), or the alternation between languages within an utterance or conversation, is common in multilingual settings but remains underrepresented in existing CS speech resources, which are typically small, domain-specific, or artificially constructed. Building on the YODAS corpus, we develop a scalable, human-in-the-loop pipeline for identifying and validating naturally occurring code-switching. The resulting dataset, which totals 313 hours and spans 7 matrix languages, provides diverse, real-world examples of spontaneous code-switched speech. We further analyze the distribution and characteristics of code-switching in the wild, examining language-pair frequencies and switching patterns, and report baseline results for spoken language identification. We hope that CS-YODAS will encourage broader and more comprehensive research on code-switched speech. Dataset link: <https://huggingface.co/datasets/byan/cs-yodas>.

Keywords: code-switching, code-switched speech, data resource construction

1. Introduction

In the current era of large-scale multilingual speech processing, models such as Whisper (Radford et al., 2023), MMS (Pratap et al., 2024), and OWSM (Peng et al., 2025) enable automatic speech recognition (ASR) and language identification (LID) across hundreds of languages. These advances are supported by massive collections of speech, providing broad coverage across languages and domains (Kahn et al., 2020; Pratap et al., 2020; Ardila et al., 2020; Wang et al., 2021; Li et al., 2023b).

Yet despite this progress, most large-scale speech resources remain *monolingual* by design: utterances are assigned a single language label, and data mining pipelines are typically optimized to exclude mixed-language content. Consequently, a core feature of multilingual communication, *code-switching*, or the alternation between languages within an utterance or conversation, is largely absent from these corpora.

Collecting code-switched speech at scale presents unique challenges. Unlike monolingual data, it requires bilingual or multilingual speakers who naturally alternate between languages, often in informal or situational contexts. Existing code-switched datasets are therefore limited in scope: they tend to be small, domain-specific, or based on elicited rather than spontaneous speech (Lyu et al., 2010; Deuchar et al., 2014; van der Westhuizen and Niesler, 2018; Chowdhury et al., 2021; Diwan et al., 2021; Hamed et al., 2022). While synthetic approaches have been proposed to fill this gap, they often fail to capture the spontaneity, prosody, and sociolinguistic nuances of naturally occurring

multilingual speech (Hussein et al., 2024; Yan et al., 2025).

In this work, we introduce CS-YODAS, a dataset of naturally occurring, in-the-wild code-switched speech mined from multilingual YouTube recordings. Our core motivation is based on the premise that better, more natural code-switched speech data will beget better, more natural code-switched speech systems. We envision that this dataset can be used as a comparison point for synthetically generated code-switching, helping practitioners identify avenues to improve the naturalness of their systems, or as training data for spoken LID systems, improving model robustness beyond what existing domain-specific and synthetically generated data can attain.

Our contributions are summarized as follows:

- A scalable, human-in-the-loop data mining procedure which enables high-precision code-switched detection on in-the-wild data
- The mined code-switched speech dataset, which consists of 313 hours of transcribed speech across 7 matrix languages
- Empirical characterizations of code-switching in the wild, along with baseline evaluations for CS-aware LID systems

Identifying code-switching in large, web-mined corpora presents a major challenge: spontaneous code-switched speech is rare and difficult to detect automatically. Off-the-shelf LID tools struggle with identifying multiple languages within utterances, particularly for noisy in-the-wild data, leading to

high false-positive rates. To address this, we design a scalable, human-in-the-loop pipeline that iteratively refines Large Language Model (LLM)-based code-switched detection using human validation. Human feedback is used to guide the LLM via in-context learning, substantially improving the precision of detected code-switched segments while maintaining scalability across web-scale speech data.

The resulting CS-YODAS dataset reflects a wide range of real-world contexts, including conversational, entertainment, and educational domains, capturing both spontaneous and semi-scripted multilingual interactions. This dataset aims to facilitate future research into both the mechanisms and modeling of natural code-switched speech.

2. Dataset Construction

2.1. Source Corpus and Motivation

CS-YODAS is derived from the YODAS corpus (Li et al., 2023b), a large-scale multilingual speech resource collected from publicly available YouTube content – by implication, CS-YODAS is under the Creative Commons license. Specifically, we use the cleaned YODAS from the OWSM v4 project (Peng et al., 2025) which consists of 166k hours of audio spanning 75 languages, with speech segments paired to automatic transcripts. While the original corpus organizes data by a single language label, preliminary inspection revealed the presence of code-switching, motivating a systematic effort to identify and extract these examples.

2.2. Overview

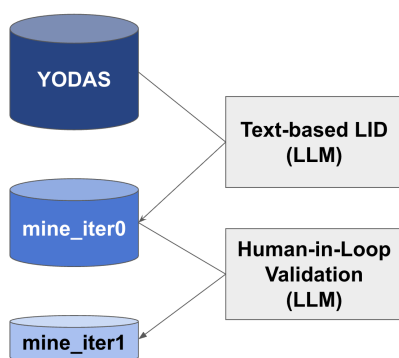


Figure 1: Data mining pipeline: we use human feedback on 700 *mine_iter0* utterances as in-context examples while prompting the LLM to validate all of *mine_iter0*, thus producing *mine_iter1*.

As shown in Figure 1, we apply a LLM-based pipeline to identify code-switched segments: first a text-based LID stage produces a set of candidate

Types	Examples
Transcript Errors	param ことないんですけど 今回ね
Proper Nouns	어려운 영어를 참 친엄마 생각해
Cognates	moment là pour la phase 2 seulement deux

Table 1: Examples of common non-CS distractors: words in **red** caused text-based LID to incorrectly predict the presence of English.

segments (*mine_iter0*) and then a human-in-the-loop validation stage produces the final CS-YODAS set (*mine_iter1*). In the remainder of this section, we describe each stage of the pipeline in detail. We also provide example prompts in §8.

2.3. Mining of Candidate Segments

Each transcript line from YODAS is passed to a multilingual LLM¹ which is prompted to infer the set of languages present in the text. The model outputs both a primary language (the dominant or matrix language) and a list of all other detected languages. Segments where two or more distinct languages are detected are retained as candidate code-switched utterances. This approach leverages the LLM’s contextual reasoning and language knowledge to capture switching phenomena that may not be easily identified through token-level LID, such as transliteration, named entities, or embedded phrases.

However, the in-the-wild nature of the source YODAS data introduces many *distractors*: noisy or imperfect transcripts can falsely mix languages, and borrowed words or named entities often resemble code-switching without representing genuine language alternation. Table 1 presents examples of the three most common types: transcription errors, proper nouns, and cognates.

2.4. Human-in-the-Loop Validation

To improve precision, we introduce a human-in-the-loop validation procedure. We sample 100 candidate segments across 7 matrix languages for human validation: Arabic, Chinese Mandarin, Czech, French, Hindi, Japanese, Russian. For each segment, annotators are asked 5 questions:

1. Is the transcript correct?
2. Does the segment contain <lang1>?

¹<https://huggingface.co/Qwen/Qwen3-14B>

Set	Samples Evaluated	Precision
mine_iter0	700	18.0%
mine_iter1	120	70.0%

Table 2: Precision of samples selected from *mine_iter0* vs *mine_iter1*, measured via agreement with human evaluations.

3. Is <lang1> the matrix language?
4. Does the segment contain <lang2>?
5. Are all <lang2> words proper nouns?

Each of these must be answered with “Yes”, “No”, or “I can’t tell”. Additionally, we solicit comments for each segment where any of Q1-4 were not answered with “Yes”.

This human feedback is then utilized via in-context learning when prompting the LLM to generate responses to the same 5 questions for all candidate segments (see §8 for example prompts). Note: we also allow the LLM to generate comments. We then filter the candidate segments using the following criteria: Q1-4 = “Yes” and Q5 = “No”. The resulting set is taken as the final CS-YODAS set (*mine_iter1*).

2.5. Evaluating the Pipeline

We conducted human evaluation on 200 candidate segments drawn from the *mine_iter0* set for each of the seven aforementioned matrix languages. Of these, 100 samples were used as in-context examples for the LLM-based validation round, while the remaining 100 were passed through the full data mining pipeline, with some ultimately appearing in the final validated set, *mine_iter1*.

As shown in Table 2, 700 candidate segments from *mine_iter0* were evaluated and only 18% were confirmed to be true instances of code-switching. Following the human-in-the-loop validation, the resulting *mine_iter1* set reached 70% precision across 120 segments identified as code-switched.

Table 3 further breaks down the effect of human-in-the-loop validation, providing the full confusion matrix across the 7 matrix languages. Hindi, Chinese, and Arabic yielded the most true positive code-switched samples at precision rates of 82.9%, 76.9%, and 55.2% respectively. Japanese and Russian yielded very few true positives; we found that samples with only proper nouns in the embedded language made up a large portion of the false negatives, indicating some annotator disagreement with our filtering methodology (namely requiring Q5 = “No”). These results highlight variation in the reliability of automatic code-switch detection across

Lang	TP	FP	FN	TN	Pre. (%)	Rec. (%)
ara	16	13	20	51	55.2	44.4
ces	1	1	1	97	50.0	50.0
cmn	20	6	7	67	76.9	74.1
fra	9	6	13	72	60.0	40.9
hin	34	7	8	51	82.9	81.9
jpn	2	3	10	50	40.0	16.7
rus	2	0	28	70	100.0	6.7
all	84	36	87	458	70.0	49.1

Table 3: Full confusion matrix breakdown of the human-in-the-loop stage of the data mining pipeline. TP = True Positive, FP = False Positive, FN = False Negative, Pre. = Precision, Rec. = Recall.

Lang	CS Duration	Total Duration	CS Rate (%)
ara	0.6	2.8	21.4
ces	0.0	0.3	13.7
cmn	0.5	4.1	12.1
fra	28.5	273.7	10.4
hin	6.3	21.2	29.7
jpn	0.3	2.0	15.0
rus	1.1	8.9	12.4
all	37.3	312.7	11.9

Table 4: Durations (hours) per matrix language: “CS Duration” refers to intra-sentential CS segments while “Total Duration” refers to contextual chunks (CS segment + 15 sec of left/right pad). CS Rate = CS Duration / Total Duration.

matrix languages, underscoring the importance of gathering language-specific human feedback.

2.6. Capturing Code-Switching Context

With CS-YODAS, we aim to facilitate the study of when and where code-switching arises in real-world communication contexts. To this end, we construct CS-YODAS in a contextual manner as follows. As described in the previous sections, we first mine a set of target segments with intra-sentential code-switching – these are typically short (<5 sec) utterances. We then extract the surrounding contexts (15 sec before and after) for these target segments and concatenate, producing a contextual chunk. In the case that multiple target segments are within 30 seconds of each other, we simply merge them into one contiguous contextual chunk.

As shown in Table 4, the total duration of contextual audio for the seven matrix languages is 312.7 hours. Of this, 37.3 hours (11.9%) correspond to intra-sententially code-switched speech. Certain languages, such as Arabic and Hindi, exhibit a higher code-switching rate.

2.7. Dataset Statistics

Dataset	Matrix Langs	Domain	Speech Type	Hours
CS-FLEURS	52	Wikipedia	Read/Synthetic	294
- read-set	14	Wikipedia	Read	17
- xtts-set	16	Wikipedia	Synthetic	221
- mms-set	45	Wikipedia	Synthetic	56
CS-YODAS	7	Youtube	Spontaneous	313

Table 5: Summary of CS-YODAS vs. CS-FLEURS.

Summary statistics are presented in Table 5. The final CS-YODAS release comprises 313 hours of speech across 7 matrix languages.

Compared to previous collections, CS-YODAS uniquely offers spontaneous code-switched speech across a wide range of languages. Earlier efforts to expand the language coverage of code-switched corpora have largely relied on synthetic generation. For example, CS-FLEURS (Yan et al., 2025) was created by first generating code-switched Wikipedia text and then recording it as read speech or synthesizing it using TTS. While such methods are valuable for scaling data collection, they fall short in capturing the natural dynamics of code-switching: for instance, when and why speakers alternate between languages in real contexts.

3. Dataset Analyses

3.1. Overview

The primary motivation of our dataset analysis is to augment the dataset with additional metadata, such as language roles, parts-of-speech, and text or audio domains. In this section, we describe our methodology for generating metadata and summarize the metadata distributions.

Please be aware that the following analyses are based on data which we estimate to be of 70% precision (as described in section §2.5). Please also refer to §6 for further discussion of the limitations of this dataset.

3.2. Language Distributions

First, let’s address the obvious question: how often did code-switching occur in-the-wild? Given that our method prioritizes precision over recall we cannot directly answer this question. Instead, we try to provide a lower-bound estimate by reporting the data mining yield rate (i.e. the number of segments identified as code-switched out of the total number of source segments) – we expect that the actual rate of code-switching in-the-wild is somewhat higher than the yield rate. As shown in Figure 2, the

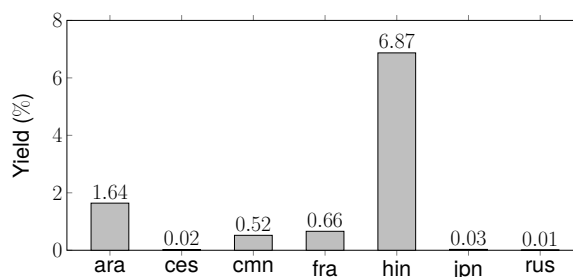


Figure 2: Yield (%) of segments identified as code-switched out of the total number of segments across matrix languages.

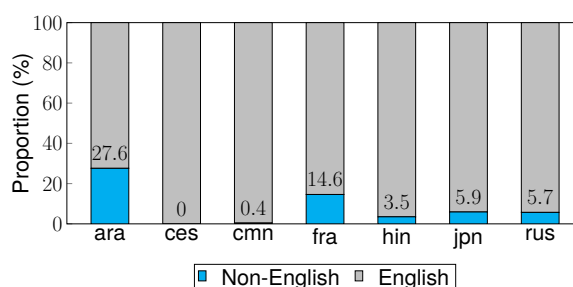


Figure 3: Proportions (%) of examples with English vs. Non-English as the embedded language across matrix languages.

yield varied greatly across matrix languages, ranging from 0.01% for Russian to 6.87% Hindi. Arabic (1.64%), French (0.66%), and Chinese (0.52%) also exhibited moderate yield rates.

Next, let’s examine the distribution of embedded languages. English is by far the most common embedded language, accounting for 85.6% of the data. Figure 3 shows the English vs. Non-English embedded language breakdown across the 7 matrix languages. Arabic (27.6%) and French (14.6%) exhibit the highest rates of code-switching with non-English languages. For Arabic, Egyptian Arabic accounted for the highest proportion of non-English code-switching. For French, Arabic accounted for the highest rate of non-English code-switching.

3.3. Lexical Distributions

Next, we examine the lexical distribution of the embedded English words to better understand the conversational purpose of code-switching. Our analysis focuses on English words to ensure reliable part-of-speech (POS) tagging. We first extract all English words and run POS tagging using spaCy. Each word is then assigned to one of four categories: *content words*, *function words*, *proper nouns*, and *discourse markers*. Content words include nouns, verbs, adjectives, and ad-

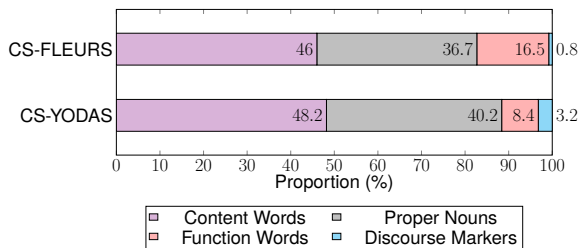


Figure 4: Proportions (%) of embedded English words across 4 categories: content words, proper nouns, function words, and discourse markers.

verbs. Proper nouns are treated as a separate category. Function words include determiners, adpositions, conjunctions, particles, and auxiliaries. Finally, discourse markers include interjections such as “yeah”, “like”, “well”, etc.

Figure 4 plots the proportions of the four lexical categories for both CS-YODAS and CS-FLEURS; the denominator here is the total count of embedded English words. Both sets contain mostly content words and proper nouns among their embedded English words. However, CS-YODAS shows a much smaller proportion of function words, nearly half that of CS-FLEURS – this suggests that the lexical pattern of the synthesized CS-FLEURS text may be unnatural. This finding aligns with linguistic theory, which suggests that function words should be least commonly used in the embedded language (Myers-Scotton, 1993). Finally, CS-YODAS includes some discourse markers, which are typical in conversational speech, whereas CS-FLEURS has none, consistent with its synthetic nature.

3.4. Domain Distributions

To further characterize the dataset, we obtain topical domain labels by feeding transcripts to a multilingual text-based domain classification model.² We then use these domain labels to estimate where code-switching tends to occur. To do so, we compare the proportional representation of each domain in CS-YODAS to that in the original YODAS corpus. For each domain, we compute the ratio between its relative frequency in CS-YODAS and its relative frequency in YODAS. A value greater than one indicates that the domain is over-represented in CS-YODAS relative to the source YODAS data, while values below one indicate under-representation. This approach normalizes for overall corpus composition and highlights domains where code-switching occurs disproportionately often.

As shown in Figure 5, informal and tech-oriented domains such as *Games*, *Internet and*

²<https://huggingface.co/nvidia/multilingual-domain-classifier>

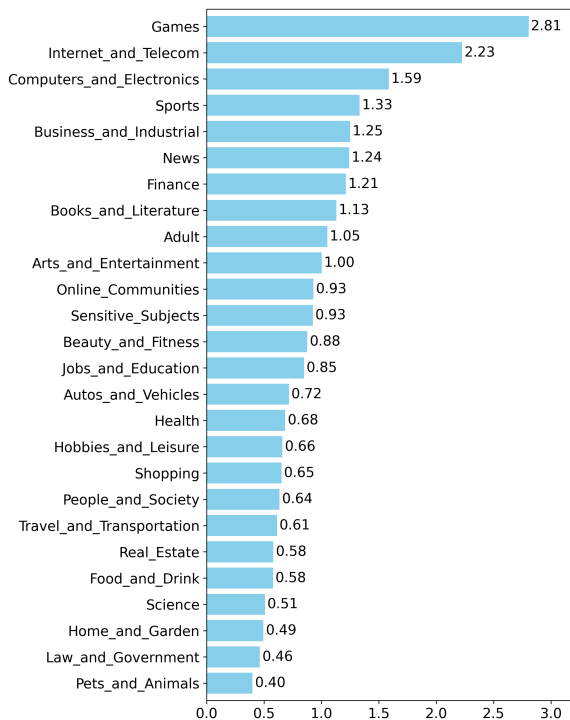


Figure 5: CS likelihood by domain (CS-YODAS proportion / YODAS proportion): values greater/less than 1 suggest higher/lower CS likelihood.

Telecom, and *Computers and Electronics* show the strongest over-representation, suggesting that code-switching frequently arises in conversational, online, and technology-related contexts. Moderately elevated ratios are also observed in *Sports*, *Business*, and *News*, where English terms or expressions are often borrowed. In contrast, more formal or topic-specific domains, such as *Law and Government*, *Health*, *Science*, and *People and Society*, show lower ratios, indicating that code-switching is less common in structured or institutional settings.

4. Baseline Experiments

4.1. Overview

Modern multilingual speech recognition and translation systems rely on accurate spoken LID both for curating large-scale training datasets (Valk and Alumäe, 2021; Peng et al., 2025) and for routing input to the language-specific modules (Radford et al., 2023; Pratap et al., 2024). However, current spoken LID systems (Jia et al., 2023; Wang et al., 2025) are almost exclusively trained under the assumption that utterances are purely monolingual, ignoring the possibility that a single utterance may contain multiple languages. Prior work has shown that this monolingual bias greatly limits the quality of downstream speech processing for code-switched

Train Set	FLEURS	CS-FLEURS							
		XTTS1	MMS	READ					
				ara-eng	cmn-eng	fra-eng	hin-eng	jpn-eng	rus-eng
w/o CS-YODAS	97.4	99.7	99.9	0	0	0	0	0	0
w/ CS-YODAS	96.3	99.5	99.8	0	0.3	51.1	19.3	0	0

Table 6: Comparison of spoken LID models with and without CS-YODAS training data. Accuracy (%) is reported on FLEURS and CS-FLEURS.

speech (Peng et al., 2023; Yan et al., 2025).

This limitation reflects a broader gap in LID research. Burchell et al. (2024) show that even in the text domain, utterance-level code-switched LID remains challenging. In the spoken domain, prior LID studies (Rangan et al., 2020; Li et al., 2023a) are limited to only a few language pairs. We argue that this gap primarily stems from a data bottleneck: large-scale multilingual speech corpora with natural code-switching and reliable language labels have been lacking.

In this work, we leverage CS-YODAS alongside the monolingual FLEURS dataset and the synthetic code-switched CS-FLEURS dataset to establish baseline LID systems that explicitly account for both monolingual and code-switched utterances. Our findings show that synthetic code-switched training data alone is insufficient to support robust generalization, and that exposure to spontaneous, naturally occurring code-switching is essential. These results highlight the importance and feasibility of curating large-scale spontaneous code-switched resources to enable more realistic spoken LID in the wild.

4.2. Experimental Setup

We compare two training data configurations: (1) simply combining the FLEURS (Conneau et al., 2023) training set with the XTTS-generated training data from CS-FLEURS and (2) further including CS-YODAS data.³ Both configurations contain 102 monolingual languages and 16 code-switched language pairs. The 16 pairs originate from CS-FLEURS, with 6 of them further supplemented by data from CS-YODAS. We refer to these settings as “w/o CS-YODAS” and “w/ CS-YODAS” respectively.

Our LID models are based on self-supervised representations, with MMS (Pratap et al., 2024) as the upstream encoder and ECAPA-TDNN (Desplanques et al., 2020) as the downstream embedding extractor. Classification is performed using the

³We only use the examples with English as the embedded language for this experiment, and exclude ces due to its limited data.

AAMSoftmax loss (Deng et al., 2019) with the sub-center enhancement (Zhao et al., 2021), following the ESPnet-SPK implementation (Jung et al., 2024). For code-switched utterances, each language pair is treated as a distinct class (i.e. “cmn” and “cmn-eng” are distinct). Training hyperparameters and optimization settings are kept consistent with the VoxLingua107-only setup described in Wang et al. (2025). The final models are selected based on the best accuracy on the FLEURS dev set and CS-YODAS XTTS1 test set. Models are evaluated on FLEURS and CS-FLEURS test sets.

4.3. Results and Discussion

As shown in Table 6, the model “w/o CS-YODAS” performs well on in-domain settings, including FLEURS and the CS-FLEURS XTTS1, as well as the out-of-domain synthetic CS-FLEURS MMS. However, the model fails across all languages in CS-FLEURS READ; the model is unable to distinguish code-switched utterances. We posit that this is due to a domain mismatch resulting from a lack of natural code-switched training data.

After incorporating the CS-YODAS training data, performance improves on the fra-eng (0% to 51.1%) and hin-eng (0% to 19.9%) subsets of the CS-FLEURS READ. This *zero-to-one* improvement demonstrates that exposure to in-the-wild code-switched speech enables models to begin generalizing to natural switching patterns beyond what synthetic data can support.

As shown in Figure 6, there is a clear trend between code-switched training data duration and code-switched LID performance: accuracy does not rise past 0 until after 5 hours of code-switched training data is available.

Our hope is that these baselines demonstrating the impact of in-the-wild code-switched training data motivates future dataset construction efforts, as our belief is that solving the data problem is critical towards building effective code-switched spoken LID systems.

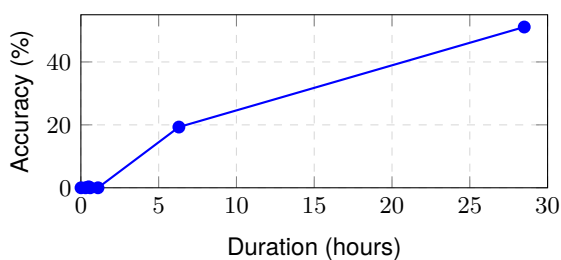


Figure 6: Accuracy (%) on CS-FLEURS READ vs. Duration (hours) of CS-YODAS training data.

5. Conclusion

In summary, our work introduces CS-YODAS, a large-scale dataset of spontaneous, naturally occurring code-switched speech. Using CS-YODAS along with the existing CS-FLEURS, we are able to show that synthetically generated code-switching has yet to thoroughly mimic the patterns observed in the wild. Further, we show that synthetic code-switched training data has yet to obviate the need for collecting spontaneous code-switched speech. Our hope is that CS-YODAS facilitates future research towards building robust spoken LID systems capable of handling all nuances of multilingual communication.

6. Limitations

While CS-YODAS represents a significant step toward large-scale, naturally occurring code-switched speech resources, several limitations remain.

First, the underlying source corpus, YODAS, is based on Creative Commons YouTube content, which inherently biases the dataset toward publicly available and broadcast-style material, such as news, educational talks, and announcements. While the dataset still captures code-switching in daily, informal contexts, we suspect that the casual setting may be underrepresented.

Second, while we implemented a human-in-the-loop validation pipeline to improve the precision of mined code-switched examples, the scale of the data precludes exhaustive manual verification. We also acknowledge other the presence of residual noise from the source corpus, such as transcription and audio segmentation errors.

Finally, reliable evaluation of code-switching requires skilled bilingual annotators. This reliance on language-specific expertise is a limiting factor against scaling coverage, and thus this release of CS-YODAS spans only 7 matrix languages.

7. Ethics Statement

This dataset is a mined subset from an already publicly released dataset, so we do not foresee any harm arising from its content. Annotators were recruited for this project on a volunteer basis and were made aware of our data mining approach and thus agreed to the use of their feedback to refine the dataset.

8. Example LLM Prompts

8.1. Overview

This section provides examples of the prompts used for text-based LID and human-in-the-loop Validation, supplementing the main description of the LLM-based data mining pipeline in §2.

8.2. Text-based LID

(System) You are performing text-based language identification. We are trying to identify code-mixed or code-switched utterances.

(User) Text: “这些地区人口稀少， often 不存在光污染的问题，你 also 能欣赏到璀璨星空。”

For the given text in triple backticks identify ALL languages that appear. There may be only a single language or multiple languages that are code-mixed together. Your final answer should list the languages in order of prevalence.

Code-mixing, or code-switching, is defined as the alternation of two languages within a single discourse, sentence, or constituent. Double check whether the text contains code-switching by reviewing word-by-word. Do not simply glance at the overall sentence and only write down the dominant language.

Format your response as a json object.

(Assistant) {"languages": ["Chinese", "English"]}

... 3 in-context examples are provided in total ...

(User) <Prompt with target segment>

(Assistant) <Text-based LID output>

Table 7: Example prompt used for text-based LID. In-context examples are depicted in blue and target segment prompt/response are depicted in violet.

Table 7 illustrates an example prompt used for the first stage of our data mining pipeline. We task the LLM with identifying all languages that appear in the segment. We use only 3 in-context examples in this stage; this choice was made in an attempt to maximize throughput.

8.3. Human-in-the-Loop Validation

(System) You are performing data validation. We are looking for code-switched speech examples by examining transcript text.

(User) Text: “因此受到他的祝福之下就代表的是权威Authority”

For the given text in triple backticks, answer the following 5 questions:

1. Is the transcript correct?
2. Does the speech contain Chinese?
3. Is Chinese the matrix language?
4. Does the speech contain English?
5. Are all English words proper nouns?

(Assistant) {"Q1": "Yes", "Q2": "Yes", "Q3": "Yes", "Q4": "Yes", "Q5": "No", "Comments": ""}

... 100 in-context examples are provided in total ...

(User) <Prompt with target segment>

(Assistant) <Human-in-the-loop Validation output>

Table 8: Example prompt used for human-in-the-loop validation. In-context examples are depicted in blue and target segment prompt/response are depicted in violet.

Table 8 illustrates an example prompt used for the second stage of our data mining pipeline. We task the LLM with replicating the human feedback pattern. We use 100 in-context examples in this stage, sourced from the human feedback on same-language samples; this choice was made in an attempt to maximize accuracy.

9. References

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 *Proc. LREC*, pages 4218–4222.

Laurie Burchell, Alexandra Birch, Robert P Thompson, and Kenneth Heafield. 2024. Code-switched language identification is harder than you think. *arXiv preprint arXiv:2402.01505*.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching Arabic ASR. In *Proc. Interspeech*, pages 2466–2470.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *Proc. IEEE SLT*, pages 798–805.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proc. CVPR*.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. Interspeech*, pages 3830–3834.

Margaret Deuchar, Peredur Davies, Jon Herring, M Carmen Parafita Couto, and Diana Carter. 2014. Building bilingual corpora. *Advances in the Study of Bilingualism*, pages 93–111.

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 *Proc. Interspeech*, pages 2446–2450.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. ArzEn-ST: A three-way speech translation corpus for code-switched Egyptian Arabic-English. In *Proc. WANLP*, pages 119–130.

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 *Proc. ICASSP*, pages 12006–12010.

Fei Jia, Nithin Rao Koluguri, Jagadeesh Balam, and Boris Ginsburg. 2023. A compact end-to-end model with local and global context for spoken language identification. In *Proc. Interspeech*, pages 5321–5325.

Jee-weon Jung, Wangyou Zhang, Jiatong Shi, Zakaria Aldeneh, Takuya Higuchi, Barry-John Theobald, Ahmed Hussen Abdelaziz, and Shinji Watanabe. 2024. ESPnet-SPK: Full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models. *arXiv preprint arXiv:2401.17230*.

- Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-Light: A benchmark for ASR with limited or no supervision. In *Proc. ICASSP*, pages 7669–7673.
- Shuyue Stella Li, Cihan Xiao, Tianjian Li, and Bismarck Odoo. 2023a. Simple yet effective code-switching language identification with multitask pre-training and transfer learning. *arXiv preprint arXiv:2305.19759*.
- Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023b. YODAS: YouTube-oriented dataset for audio and speech. In *Proc. IEEE ASRU*, pages 1–8.
- Dau-Cheng Lyu, Tien Ping Tan, Engsiong Chng, and Haizhou Li. 2010. SEAME: a mandarin-english code-switching speech corpus in south-east asia. In *Proc. Interspeech*, pages 1986–1989.
- C. Myers-Scotton. 1993. *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press.
- Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath. 2023. Prompting the hidden talent of web-scale speech models for zero-shot task generalization. In *Proc. Interspeech*, pages 396–400.
- Yifan Peng, Shakeel Muhammad, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. 2025. OWSM v4: Improving open Whisper-style speech models via data scaling and cleaning. *Proc. Interspeech*, pages 2225–2229.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, pages 1–52.
- Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Proc. Interspeech*, pages 2757–2761.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, pages 28492–28518.
- Pradeep Rangan, Sundeep Teki, and Hemant Misra. 2020. Exploiting spectral augmentation for code-switched spoken language identification. *arXiv preprint arXiv:2010.07130*.
- Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a dataset for spoken language recognition. In *Proc. IEEE SLT*, pages 652–658.
- Ewald van der Westhuizen and Thomas Niesler. 2018. A first south African corpus of multilingual code-switched soap opera speech. In *Proc. LREC*.
- Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proc. ACL-IJCNLP*.
- Qingzheng Wang, Hye-jin Shim, Jiancheng Sun, and Shinji Watanabe. 2025. Geolocation-aware robust spoken language identification. In *Proc. IEEE ASRU*.
- Brian Yan, Injy Hamed, Shuichiro Shimizu, Vasista Sai Lodagala, William Chen, Olga Iakovenko, Bashar Talafha, Amir Hussein, Alexander Polok, Calvin Chang, et al. 2025. CS-FLEURS: A massively multilingual and code-switched speech dataset. In *Proc. Interspeech*, pages 743–747.
- Miao Zhao, Yufeng Ma, Min Liu, and Minqiang Xu. 2021. The SpeakIn system for VoxCeleb speaker recognition challenge 2021. *arXiv preprint arXiv:2109.01989*.